Lesson 6: Normal and Poisson distributions

TB sections 3.3-3.4

Meike Niederhausen and Nicky Wakim

2024-10-16

Learning Objectives

1. Understand how probability distributions extend to continuous distributions
2. Calculate probabilities for specific events using a Normal distribution
3. Apply the Normal distribution to approximate probabilities for binomial events
4. Calculate probabilities for different events using a Poisson distribution

Where are we?

Data
- Collecting data
- Categorical vs. Numeric
- Summary statistics
- Data visualization

Probability
- Probability rules
- Independence, conditional
- Random variables and probability distributions
- Linear combinations
- Binomial, Normal, and Poisson

Sampling Variability, and Statistical Inference
- Sampling distributions
- Central Limit Theorem
- Confidence Intervals
- Hypothesis tests

Inference for continuous data/outcomes
- One sample t-test
- 3+ independent samples
- Simple linear regression / correlation
- 2 sample tests: paired and independent
- Power and sample size
- Non-parametric tests

Inference for categorical data/outcomes
- One proportion test
- Fisher's exact test
- Non-parametric tests
- Chi-squared test
- 2 proportion test
- Power and sample size

R
- Basics
- Reproducibility
- Quarto
- Packages
- Data visualization
- Data wrangling
- R Projects
- …

Learning Objectives

1. Understand how probability distributions extend to continuous distributions
2. Calculate probabilities for specific events using a Normal distribution
3. Apply the Normal distribution to approximate probabilities for binomial events
4. Calculate probabilities for different events using a Poisson distribution

Last time: Discrete vs. continuous random variables

- Probability distributions are usually either **discrete** or **continuous**, depending on whether the random variable is discrete or continuous.

Discrete random variable

A **discrete** r.v. X takes on a finite number of values or countably infinite number of possible values.

Continuous random variable

A **continuous** r.v. X can take on any real value in an interval of values or unions of intervals.

Think:

- Number of heads in a set of coin tosses
- Number of people who have had chicken pox in a random sample

- Binomial and Bernoulli distributions are discrete

Think:

- Height in a population
- Blood pressure in a population

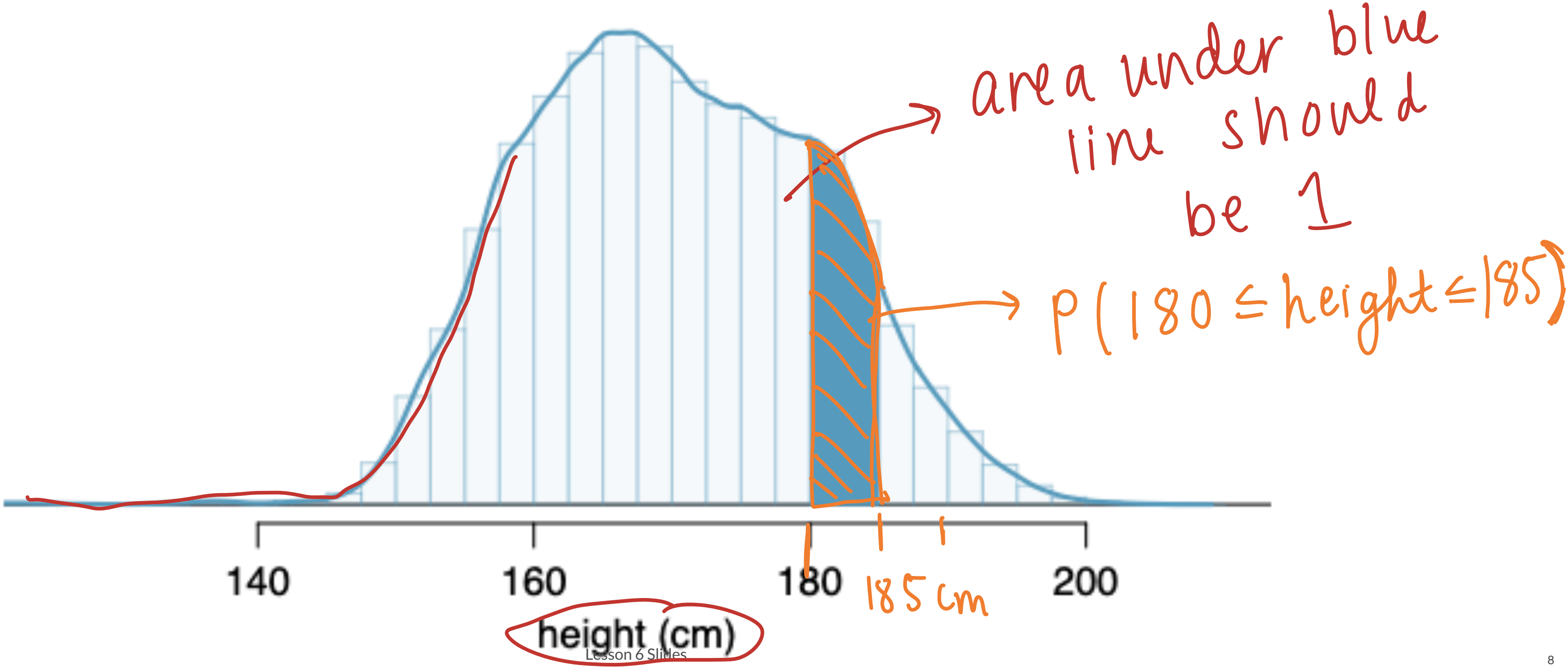
hours slept

7hrs 7.001hr ~~7~~.00001 hrs

Probabilities for continuous distributions (1/2)

Two important features of continuous distributions:

- The total area under the density curve is 1.
- The probability that a variable has a value within a specified interval is the area under the curve over that interval.



Probabilities for continuous distributions (2/2)

When working with continuous random variables, probability is found for **intervals of values** rather than **individual values**.

- The probability that a continuous r.v. X takes on any single individual value is 0
  - That is, $P(X = x) = 0$.
- Thus, $P(a < X < b)$ is equivalent to $P(a \leq X \leq b)$

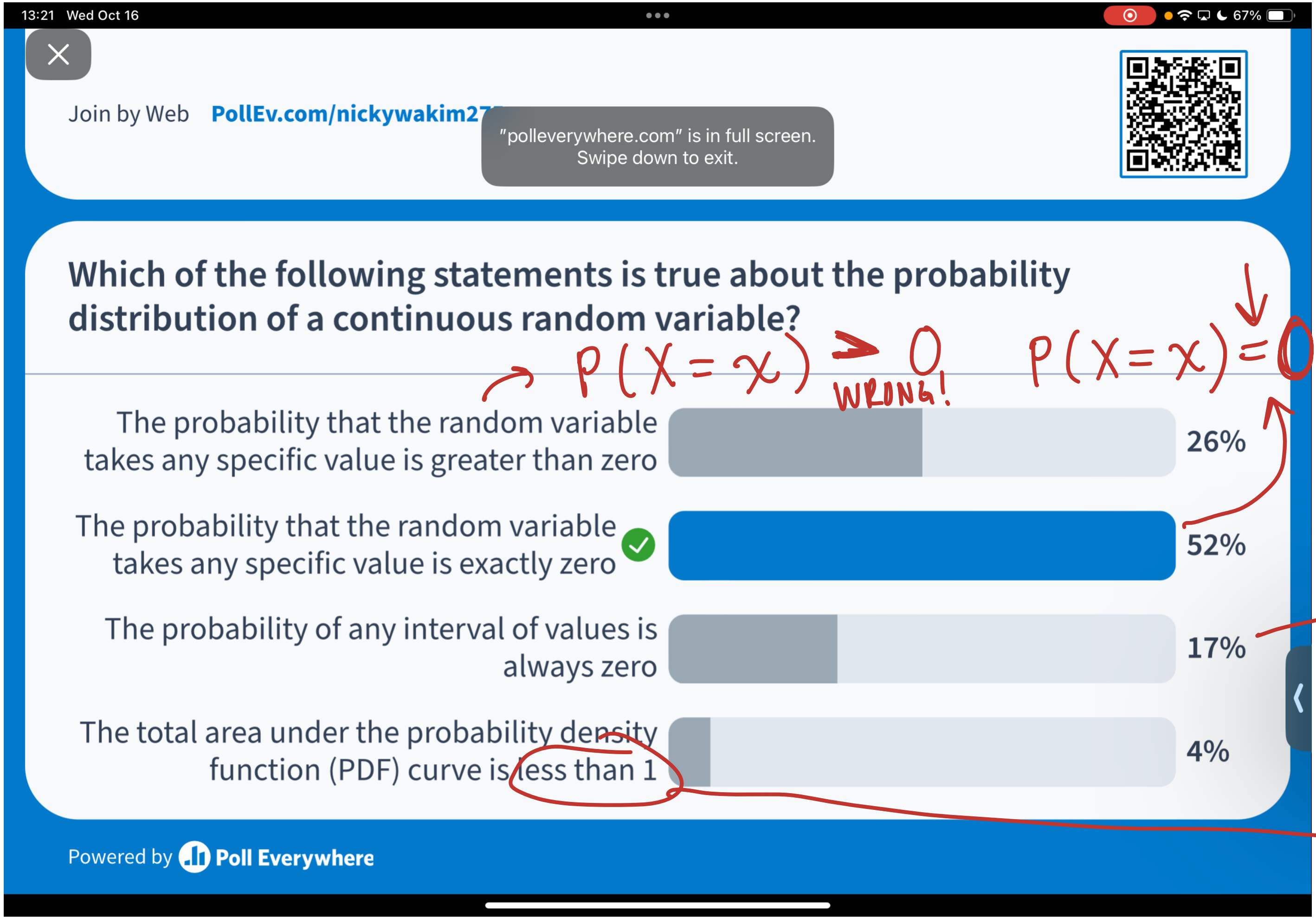
infinite possible values

$$\frac{1 \text{ value}}{\infty \text{ possible vals}} = 0$$

$$P(X = 7.01) = 0$$

$$0 \leq P(6.99 \leq X \leq 7.01) \leq 1$$

Poll Everywhere Question 1



WRONG $P(a \le X \le b) = 0$

RIGHT: $0 \le P(a \le X \le b) \le 1$

total probability <u>equals</u> 1

Learning Objectives

1. Understand how probability distributions extend to continuous distributions
2. Calculate probabilities for specific events using a Normal distribution
3. Apply the Normal distribution to approximate probabilities for binomial events
4. Calculate probabilities for different events using a Poisson distribution

Normal distribution

Continuous!

- A random variable X is modeled with a normal distribution if:
  - **Shape:** symmetric, unimodal bell curve
  - **Center:** mean μ
  - **Spread (variability):** standard deviation σ

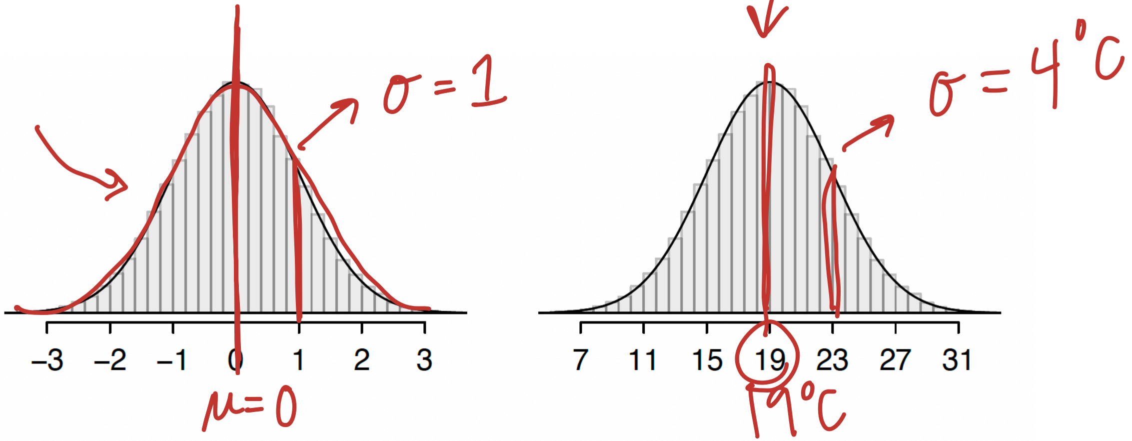


Figure 3.5: Both curves represent the normal distribution; however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

- Shorthand for a random variable, X, that has a Normal distribution:

$$X \sim \text{Normal}(\mu, \sigma)$$

"distributed by"

- **Example:** We recorded the high temperature in the past 100 years for today. The mean high is 19°C (66.2°F)

Standard Normal distribution (1/2)

$X \sim \text{Norm}(\mu, \sigma)$

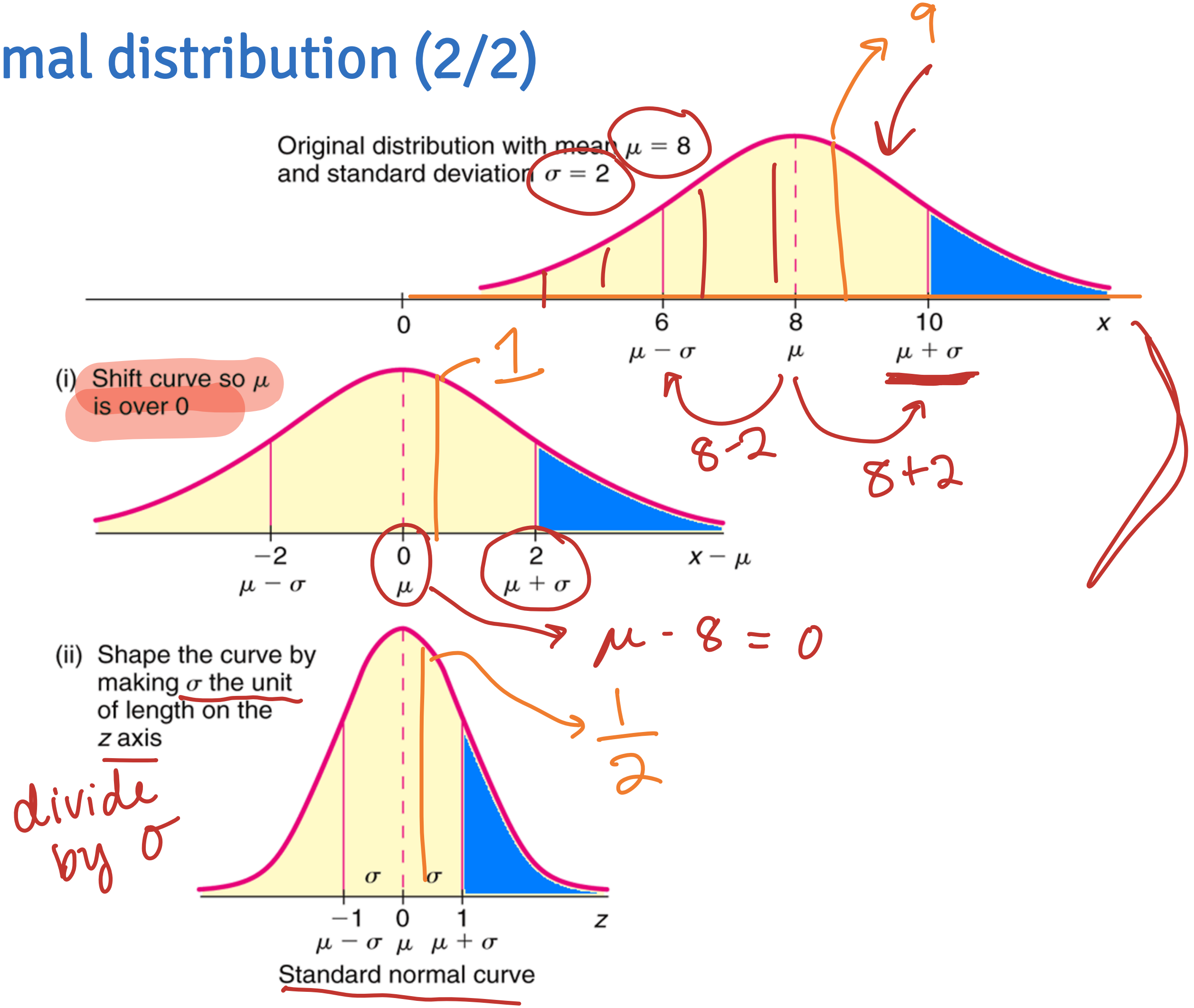
- A *standard normal* distribution is defined as a normal distribution with mean 0 and variance 1. It is often denoted as $Z \sim N(0, 1)$.
- Any normal random variable X can be transformed into a standard normal random variable Z.

standardization

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

- The Z-score of an observation quantifies how far the observation is from the mean, in units of standard deviation(s).
- For example, if an observation has Z-score $z = 3.4$, then the observation is 3.4 standard deviations above the mean.

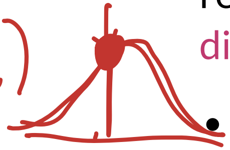
Standard Normal distribution (2/2)



Transformation from general normal X to standard normal Z

Normal distribution: R commands

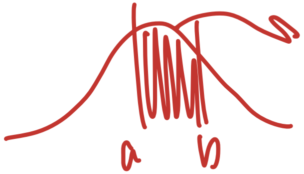
R commands with their input and output:

R code	What does it return?
`rnorm()`	returns sample of random variables with specified normal distribution
`dnorm()` $f_X(x)$ 	returns value of probability density at certain point of the normal distribution • Not typically used bc this is a continuous distributon
`pnorm()` $P(X \leq x)$	returns cumulative probability of getting certain point (or less) of the normal distribution
`qnorm()`	returns z-score corresponding to desired quantile or x-value

Calculating probabilities from a Normal distribution

Four ~~Three~~ ways to calculate probabilities from a normal distribution:

1. ~~Calculus~~ **(not for us!)**



$\int_a^b$

2. Normal probability table

z-score table

- The textbook has a normal probability table in Appendix B.1, which is included as the next two pages
- Not required for this class

3. R commands

change

$\leq$

- $P(Z \leq q) =$ `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)`

FALSE : $P(Z > q)$

4. Random online calculators

- Like this one: https://onlinestatbook.com/2/calculators/normal_dist.html

Example: Calculating probabilities from a Normal distribution (1/5)

Example: Calculating standard normal probabilities practice

Let Z be a standard normal random variable, $Z \sim N(\mu = 0, \sigma = 1)$. Calculate the following probabilities. Include sketches of the normal curves with the probability areas shaded in.

1. $\mathbb{P}(Z < 2.67)$
2. $\mathbb{P}(Z > -0.37)$
3. $\mathbb{P}(-2.18 < Z < 2.46)$
4. $\mathbb{P}(Z = 1.53)$

Example: Calculating probabilities from a Normal distribution (2/5)

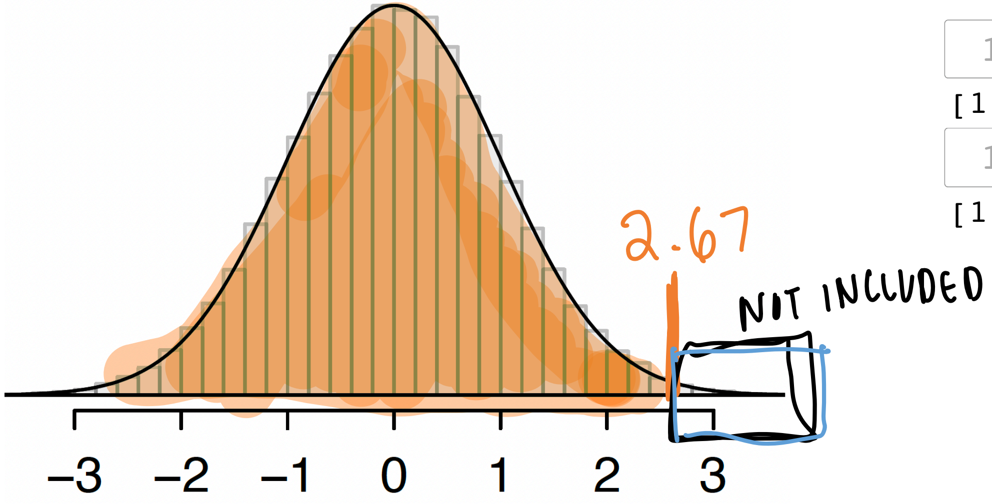
Example: Calculating standard normal probabilities practice

Let Z be a standard normal random variable, $Z \sim N(\mu = 0, \sigma = 1)$. Calculate the following probabilities. Include sketches of the normal curves with the probability areas shaded in.

1. $\mathbb{P}(Z < 2.67)$

P(Z < 2.67)

1. Draw on standard Normal curve:



2. Calculate probability:

```
pnorm(q = 2.67, mean = 0, sd = 1)
[1] 0.9962074
```

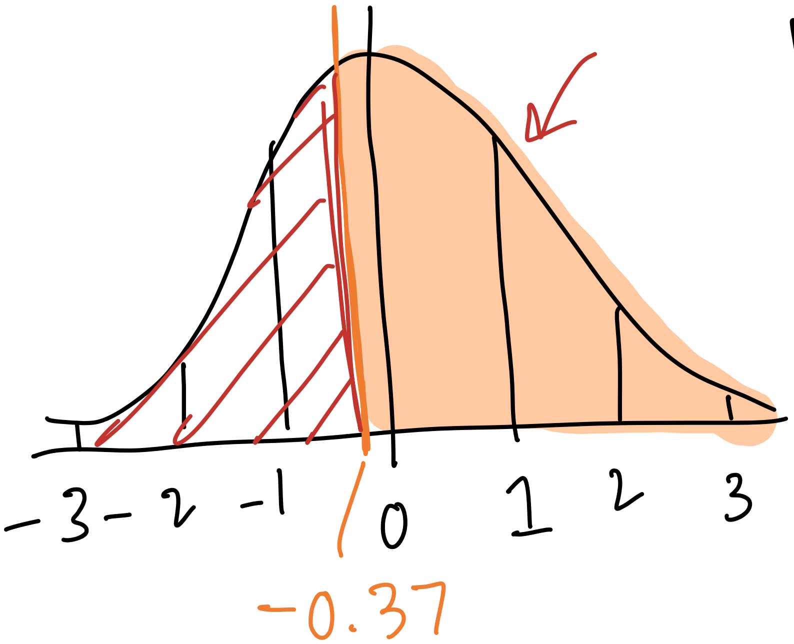
```
pnorm(q = 2.67)
[1] 0.9962074
```

Example: Calculating probabilities from a Normal distribution (3/5)

Example: Calculating standard normal probabilities practice

Let Z be a standard normal random variable, $Z \sim N(\mu = 0, \sigma = 1)$. Calculate the following probabilities. Include sketches of the normal curves with the probability areas shaded in.

2. $\mathbb{P}(Z > -0.37)$



$P(Z > -0.37)$

pnorm(q = - 0.37, lower.tail = FALSE)

0.644

$1 - \underbrace{\text{pnorm}(q = -0.37)}$

lower.tail = T

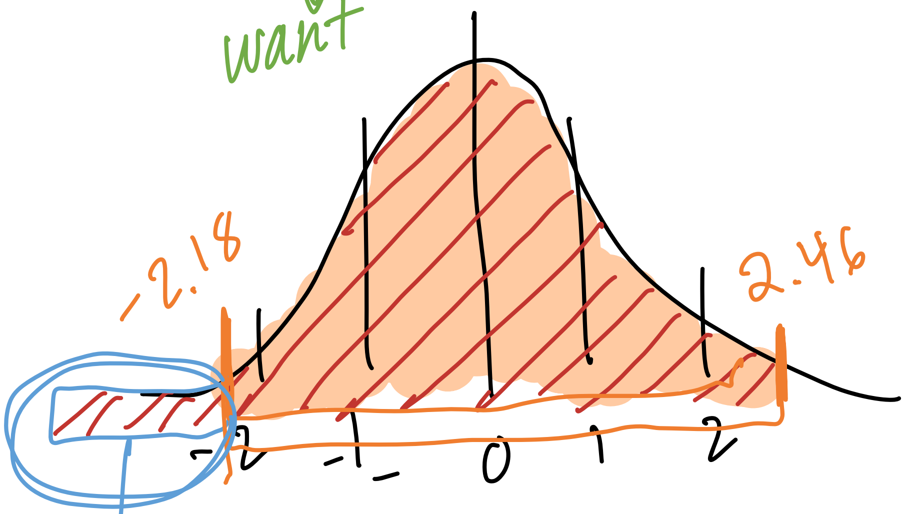
Example: Calculating probabilities from a Normal distribution (4/5)

Example: Calculating standard normal probabilities practice

Let Z be a standard normal random variable, $Z \sim N(\mu = 0, \sigma = 1)$. Calculate the following probabilities. Include sketches of the normal curves with the probability areas shaded in.

3. $\mathbb{P}(-2.18 < Z < 2.46)$

want



remove this area

$$P(-2.18 < Z < 2.46)$$

$$= P(Z < 2.46) - P(Z < -2.18)$$

in R:

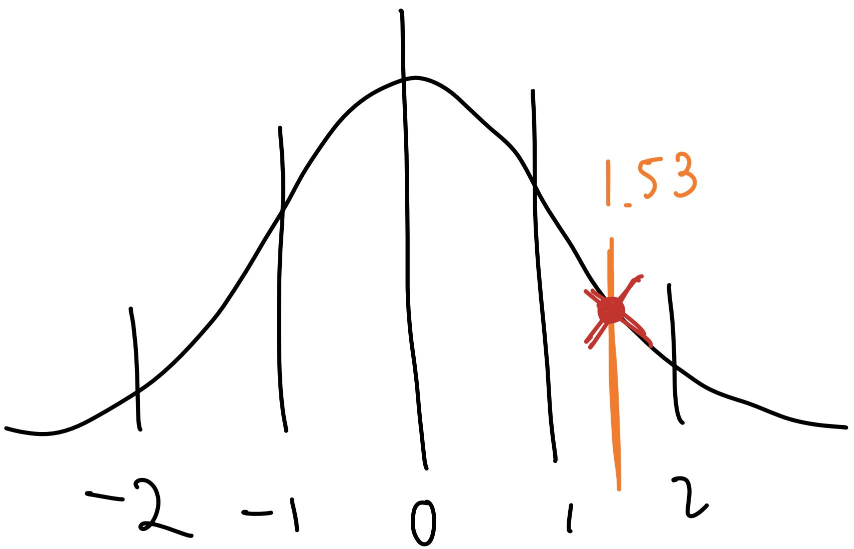
```
pnorm(q = 2.46) - pnorm(q=-2.18)
```

Example: Calculating probabilities from a Normal distribution (5/5)

Example: Calculating standard normal probabilities practice

Let Z be a standard normal random variable, $Z \sim N(\mu = 0, \sigma = 1)$. Calculate the following probabilities. Include sketches of the normal curves with the probability areas shaded in.

4. $\mathbb{P}(Z = 1.53)$



$P(X = x) = 0$ for cont dist'ns

$P(Z = 1.53) = 0$

Using Normal distribution in word problems

Example: Diastolic blood pressure (DBP)

Suppose the distribution of diastolic blood pressure (DBP) in 35- to 44-year old men is normally distributed with mean 80 mm Hg and variance 144 mm Hg.^{2} var units^2 sd units

1. Mild hypertension is when the DBP is between 90 and 99 mm Hg. What proportion of this population has mild hypertension?
2. What is the 10^{th} percentile of the DBP distribution?
3. What is the 95^{th} percentile of the DBP distribution?

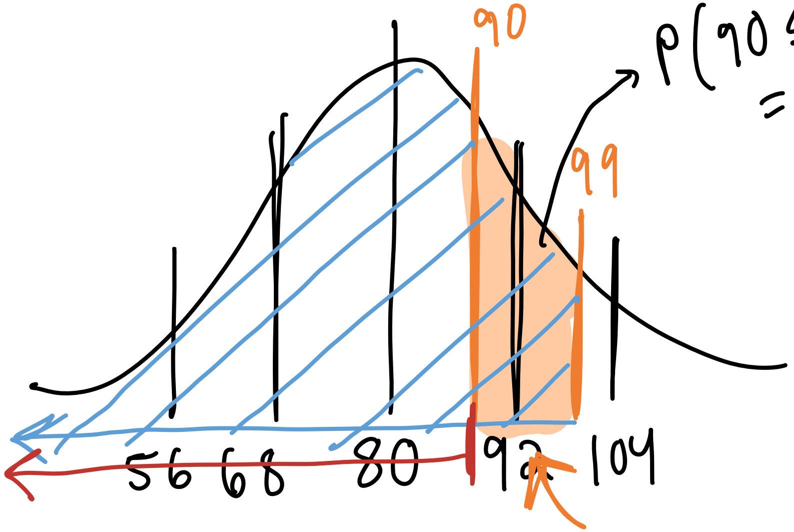
Using Normal distribution in word problems

Example: Diastolic blood pressure (DBP)

Suppose the distribution of diastolic blood pressure (DBP) in 35- to 44-year old men is normally distributed with mean 80 mm Hg and variance 144 mm Hg^2. $\cancel{\sqrt{144}} =$ $sd = \sqrt{var} = \sqrt{144} = 12$

1. Mild hypertension is when the DBP is between 90 and 99 mm Hg. What proportion of this population has mild hypertension?

- Draw on a normal curve:



$P(90 \le X \le 99)$ = orange area

- Compute in R:

$P(X \le 99)$

```
pnorm(q = 99, mean = 80,
      sd = sqrt(144)) -
  pnorm(q = 90, mean = 80,
        sd = sqrt(144))
```

$P(X \le 90)$

```
[1] 0.1456556
```

Using Normal distribution in word problems

Example: Diastolic blood pressure (DBP)

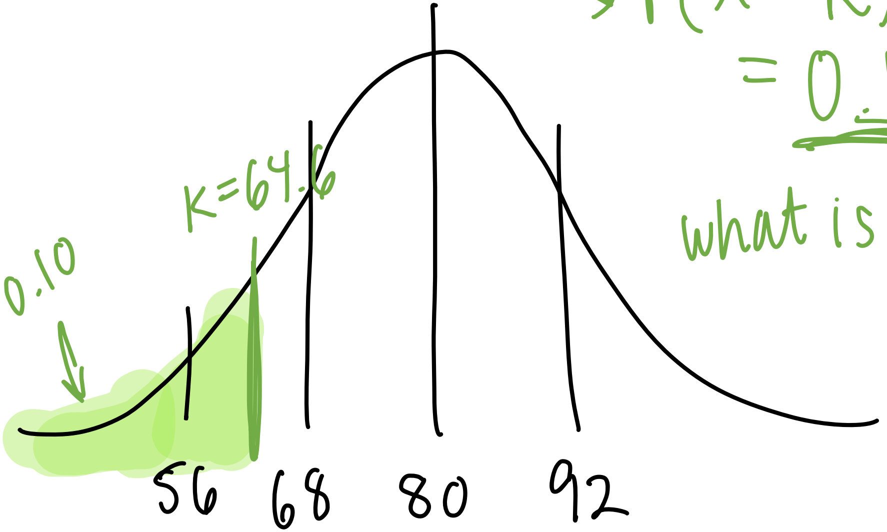
Suppose the distribution of diastolic blood pressure (DBP) in 35- to 44-year old men is normally distributed with mean 80 mm Hg and variance 144 mm Hg.

2. What is the 10^{th} percentile of the DBP distribution?

- Draw on a normal curve:

$P(X \le k) = 0.10$

what is k?



- Compute in R:

```
qnorm(p = 0.10,
      mean = 80,
      sd = sqrt(144))
```

```
[1] 64.62138
```

$P(X \le 64.62) = 0.10$

Using Normal distribution in word problems

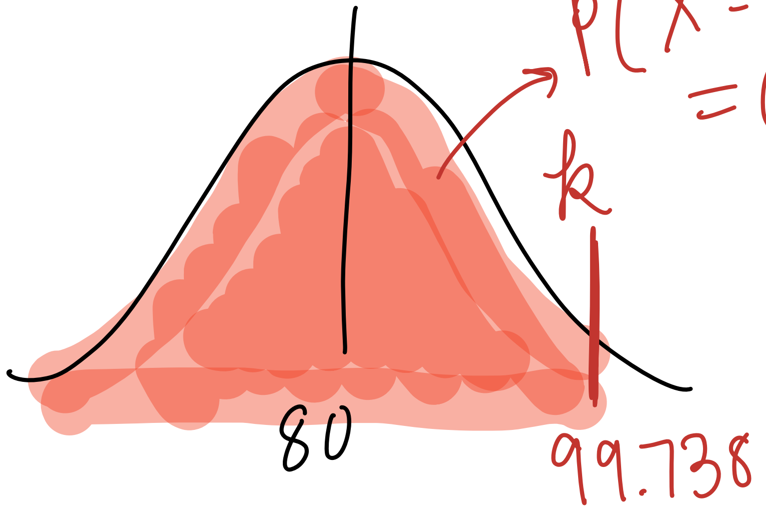
Example: Diastolic blood pressure (DBP)

Suppose the distribution of diastolic blood pressure (DBP) in 35- to 44-year old men is normally distributed with mean 80 mm Hg and variance 144 mm Hg.

3. What is the 95^{th} percentile of the DBP distribution?

- Draw on a normal curve:

$P(X \le k) = 0.95$



- Compute in R:

```
qnorm(p = 0.95,
      mean = 80,
      sd = sqrt(144))
```

```
[1] 99.73824
```

Learning Objectives

1. Understand how probability distributions extend to continuous distributions
2. Calculate probabilities for specific events using a Normal distribution
3. Apply the Normal distribution to approximate probabilities for binomial events
4. Calculate probabilities for different events using a Poisson distribution

Normal Approximation of the Binomial Distribution

- Recall that a binomial random variable X counts the total number of successes in n independent trials, each with probability p of a success.

- Probability function for $x = 0, 1, \ldots, n$:

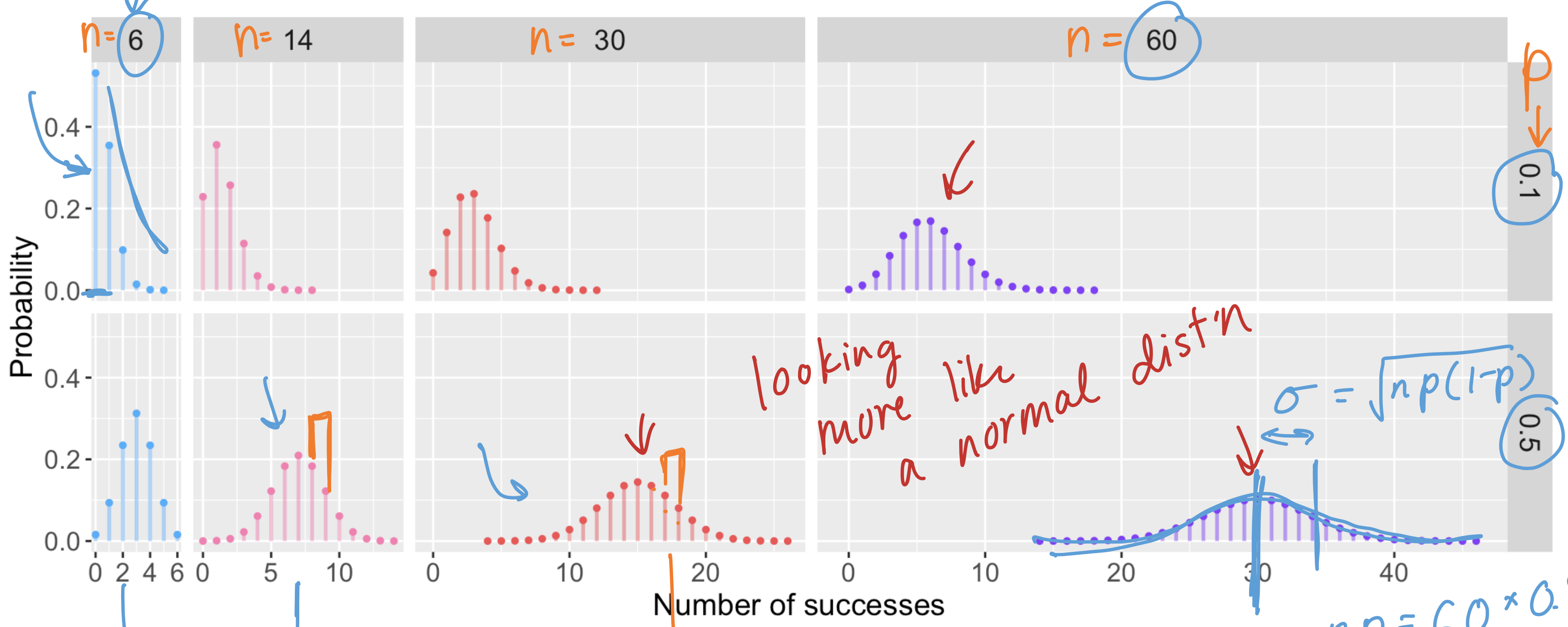
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

- Tedious to compute for large number of trails (n), although doable with software like R

- As n gets big though, the distribution shape of a binomial r.v. gets more and more symmetric, and can be approximated by a normal distribution
- Pretty good video behind the intuition of this (Watch 00:00 - 05:40)

We can look at a plot of Binomial distributions

- Binomial distributions for different n (columns) and p (rows) prob of success



$6 \times 0.5 = 3$

$14 \times 0.5 = 7$

17 - 18

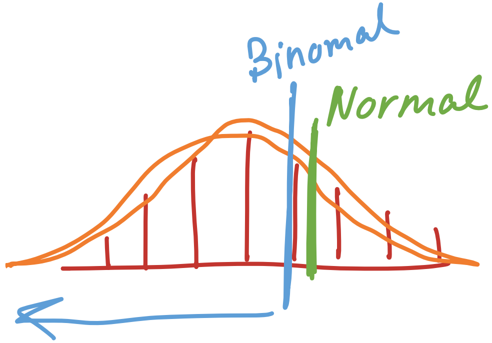
$\mu = np = 60 \times 0.5$

Normal Approximation of the Binomial Distribution

- Also known as: **Sampling distribution of $\hat{p}$**
- If $X \sim \text{Binomial}(n, p)$ and $np > 10$ and $nq = n(1-p) > 10$
  - Ensures sample size (n) is moderately large and the p is not too close to 0 or 1
  - Other resources use other criteria (like $npq > 5$ or $np > 5$)

p is prob of success for binomial

$n(p)(1-p) > 5$



- THEN approximately

$$X \sim \text{Normal}\left(\mu_X = np, \sigma_X = \sqrt{np(1-p)}\right)$$

- **Continuity Correction:** Applied to account for the fact that the binomial distribution is discrete, while the normal distribution is continuous

gaps in counts for binomial

  - Adjust the binomial value (# of successes) by ±0.5 before calculating the normal probability.
  - For $P(X \leq k)$ (Binomial), you would instead calculate $P(X \leq k + 0.5)$ (Normal approx)
  - For $P(X \geq k)$ (Binomial), you would instead calculate $P(X \leq k - 0.5)$ (Normal approx)

$P(X \leq 17)$

$P(X \leq 17 + 0.5) = P(X \leq 17.5)$

Example: Normal approximation or Binomial distribution (1/2)

Example: Vaccinated people testing positive for Covid-19 (revisited)

About 25% of people that test positive for Covid-19 are vaccinated for it. Suppose 100 people have tested positive for Covid-19 (independently of each other). Let X denote the number of people that are vaccinated among the 100 that tested positive. What is the probability that fewer than 20 of the people that tested positive are vaccinated?

1. Calculate exact probability.
2. Calculate approximate probability.

$p = 0.25, n = 100$, we want $P(X < 20)$

1. Exact probability = Binomial distribution

$$\rightarrow X \sim \text{Binomial}(n = 100, p = 0.25)$$

$$P(X < 20) = P(X \leq 19) = \sum_{j=0}^{19} P(X = j)$$

```r
pbinom(q = 19, size = 100, prob = 0.25)
```

```
[1] 0.09953041
```

25% vaxx $p = 0.25$

$n = 100$ $X \sim \text{Binom}(n = 100, p = 0.25)$

Example: Normal approximation or Binomial distribution (2/2)

Example: Vaccinated people testing positive for Covid-19 (revisited)

About 25% of people that test positive for Covid-19 are vaccinated for it. Suppose 100 people have tested positive for Covid-19 (independently of each other). Let X denote the number of people that are vaccinated among the 100 that tested positive. What is the probability that fewer than 20 of the people that tested positive are vaccinated?

1. Calculate exact probability.

2. Calculate approximate probability.

need to show we can do approx

$np \overset{?}{\geq} 10$, $n(1-p) > 10$

$100 \cdot 0.25 \overset{?}{\geq} 10$ $\quad$ $100 \cdot 0.75 \overset{?}{\geq} 10$ → can use Normal approx

$25 > 10$ $\quad$ $75 > 10$

$p = 0.25, n = 100$, we want $P(X < 20)$

2. Approximate probability = Normal distribution

- Mean = $\mu = np = 0.25 \cdot 100 = 25$
- SD = $\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0.25 \cdot (1 - 0.25)} = 4.33$

approx binomial w/ $X \sim \text{Normal}(\mu = 25, \sigma = 4.33)$

- Use **continuity correction**: Instead of calculating $P(X \leq 19)$, we calculate $P(X \leq 19.5)$

```
pnorm(q = 19.5, mean = 25,
      sd = sqrt( 100*0.25*0.75 ))
```

```
[1] 0.1020119
```

0.102 vs. 0.0995

Some resources for the normal distribution

- Page on R commands

Learning Objectives

1. Understand how probability distributions extend to continuous distributions
2. Calculate probabilities for specific events using a Normal distribution
3. Apply the Normal distribution to approximate probabilities for binomial events
4. Calculate probabilities for different events using a Poisson distribution

Introduction to the Poisson distribution

- The Poisson distribution is often used to model **count data (# of successes)**, especially for **rare events**
  - It is a **discrete distribution!**

- It is used most often in settings where **events happen at a rate** λ per unit of population and per unit time

- **Example:** historical records of hospitalizations in New York City indicate that an average of 4.4 people are hospitalized each day for an acute myocardial infarction (AMI)
  - We can plot the distribution of hospitalizations on each day

$\lambda = 4.4$ hospl day

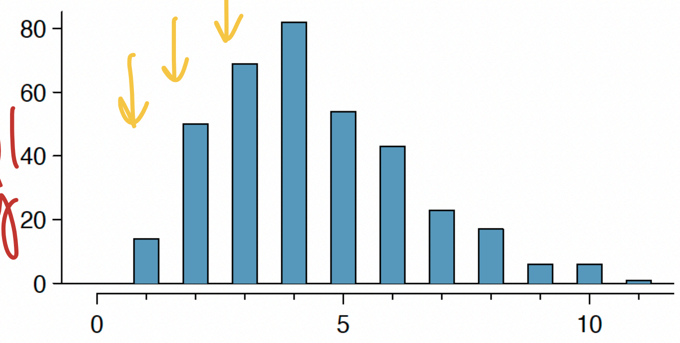


Figure 3.20: A histogram of the number of people hospitalized for an AMI on 365 days for NYC, as simulated from a Poisson distribution with mean 4.4.

Poisson distribution

- Suppose events occur over time in such a way that
  1. The probability an event occurs in an interval is proportional to the length of the interval.
  2. Events occur independently at a rate λ per unit of time.
- Then the probability of exactly x events in one unit of time is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- For the Poisson distribution modeling the number of events in one unit of time:
  - The mean is λ.
  - The standard deviation is $\sqrt{\lambda}$.
- Shorthand for a random variable, X, that has a Poisson distribution:

$$X \sim \text{Pois}(\lambda)$$

prob hosp in 1 min
vs.
prob hosp in
4 hrs

k = # of observed events
that we're interested in

same

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

$x = 0, 1, 2, \ldots$

Poisson distribution: R commands

R commands with their input and output:

R code		What does it return?
`rpois()`		returns sample of random variables with specified Poisson distribution
`dpois()`	$P(X=k)$	returns value of probability density at certain point of the Poisson distribution
`ppois()`	$P(X\leq k)$	returns cumulative probability of getting certain point (or less) of the Poisson distribution
`qpois()`		returns number of cases corresponding | to desired quantile

↳ 90th percent : $\rightarrow k$

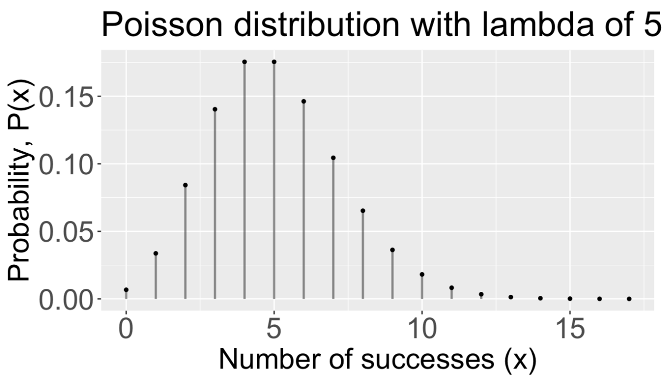
Example: probabilities from a Poisson distribution (1/4)

Typhoid fever

Suppose there are on average 5 deaths per year from typhoid fever over a 1-year period.

LOOK @

1. What is the probability of 3 deaths in a year?
2. What is the probability of 2 deaths in 0.5 years?
3. What is the probability of more than 12 deaths in 2 years?



Example: probabilities from a Poisson distribution (2/4)

Typhoid fever

Suppose there are on average 5 deaths per year from typhoid fever over a 1-year period.

1. What is the probability of 3 deaths in a year?

$\lambda = \frac{5 \text{ deaths}}{1 \text{ year}}$

- $\lambda = 5$ and we want $P(X = 3)$

$\lambda = 5$

$$P(X = 3) = \frac{e^{-5}5^3}{3!} = 0.1404$$

```
dpois(x = 3, lambda = 5)
```

```
[1] 0.1403739
```

$P(X=3) = 0.14$

Example: probabilities from a Poisson distribution (3/4)

Typhoid fever

Suppose there are on average 5 deaths per year from typhoid fever over a 1-year period.

2. What is the probability of 2 deaths in 0.5 years?

$\lambda =?$ and we want $P(X=2)$

let X = # death in 0.5 years → need $\lambda = \frac{\text{deaths}}{0.5\text{yrs}}$

- $\lambda = 5$ was the rate for one year. When we want the rate for half year, we need to calculate a new λ:
  - $\lambda = \frac{5 \text{ deaths}}{\cancel{1 \text{ year}}} \cdot \frac{\cancel{1 \text{ year}}}{2 \text{ half-years}} = \frac{2.5 \text{ deaths}}{1 \text{ half-year}}$

$\frac{5 \text{ deaths}}{2 \text{ half-yrs}}$

$\lambda = 2.5$

$$P(X=2) = \frac{e^{-2.5}2.5^2}{2!} = 0.0.2565$$

same

k

```
dpois(x = 2, lambda = 2.5)
```

```
[1] 0.2565156
```

$P(X=4)$ w/ $\lambda = 5$

Example: probabilities from a Poisson distribution (4/4)

Typhoid fever

Suppose there are on average 5 deaths per year from typhoid fever over a 1-year period.

3. What is the probability of more than 12 deaths in 2 years?

change λ

$\lambda = ?$ and we want $P(X > 12)$

- Need to calculate a new λ for 2 years: $\lambda = \frac{5 \text{ deaths}}{1 \text{ year}} \cdot \frac{2 \text{ years}}{1 \text{ two-years}} = \frac{10 \text{ deaths}}{1 \text{ two-year}} = \frac{10 \text{ deaths}}{\cancel{\text{two}}\ 2 \text{ years}} = 10$

$P(A) = 1 - P(A^c)$

$P(X \leq 12) = P(X=1) + P(X=2) + \cdots P(X=12)$

$$P(X > 12) = 1 - P(X \leq 12) = 1 - \boxed{\sum_{k=0}^{12} \frac{e^{-10} 10^k}{k!}} = 0.2084$$

```r
1 - ppois(q = 12, lambda = 10)
```

```
[1] 0.2084435
```

$1 - P(X \leq 12)$

```r
ppois(q = 12, lambda = 10,
      lower.tail = F)
```

```
[1] 0.2084435
```

$P(X > 12)$

Poisson approximation of binomial distribution

typically use Poisson approx for rare events that are Binomally dist.

- Poisson distribution can be used to approximate binomial distribution when n is large and p is small
  - When Normal approximation does not work
- Binomial distributions for different n (columns) and p (rows)