INTRODUCTION TO DATA & STATISTICAL INFERENCE

Stat 218: Applied Statistics for the Life Sciences

California Polytechnic State University - San Luis Obispo

Tuesday, OCTOBER 11, 2022

Today we will...

- Statistical Inference Recap
- Big Picture on Simple Linear Regression
- Activity 4: Diving Penguins
  - Checked for engagement and completion in class Thursday

STATISTICAL INFERENCE RECAP

Big Idea: Hypothesis testing gives us the framework to formally use statistical methods to draw inferences about our entire population from our collected data.

1. Research question
   - Set up your context
   - Write your hypotheses/claims
2. Find evidence against your claims
   - p-values
   - confidence intervals (also can just help us describe our data)
3. Make a decision about your claim
4. Interpret your results in context of the research question

SETUP

Setup

- Observational Units
- Variable of Interest (and type)
- Population of Interest
- Parameter (has a symbol - last week μ (Mu) for the true population mean)
    - think of this as a "word formula" (summary statistic + variable of interest + population)

Hypotheses

- (H_O) Null: "innocent", "nothing is going on", "status quo" (= Null Value)
- (H_A) Alternative: "guilty", "something is going on", "what we are trying to prove" (<, >, $\neq$)

STATISTICALLY SIGNIFICANT

Let's find evidence!

1. **Simulation**
    - Bootstrapping: repeated sampling with replacement
    - Randomization: we will do this today (we actually did this on day 1 of class!)
2. **Theory/Math**
    - We know certain properties of these distributions

Distributions

- Observed Data
- Sampling Distribution (Either from Bootstrapping or from Theory)
- Null Distribution (Either from Randomization or from Theory)

STATISTICAL SIGNIFICANCE

How strong is our evidence? How much evidence do we need?

α-threshold	Confidence Level
0.10	90%
0.05	95%
0.01	99%

We set our **significance level** at the beginning of a study.

P-VALUES

This is the probability of seeing your observed summary statistic (from your data) IF the null is true (aka assuming "nothing is going on").

Compare this to our null distribution (either randomization or theory)

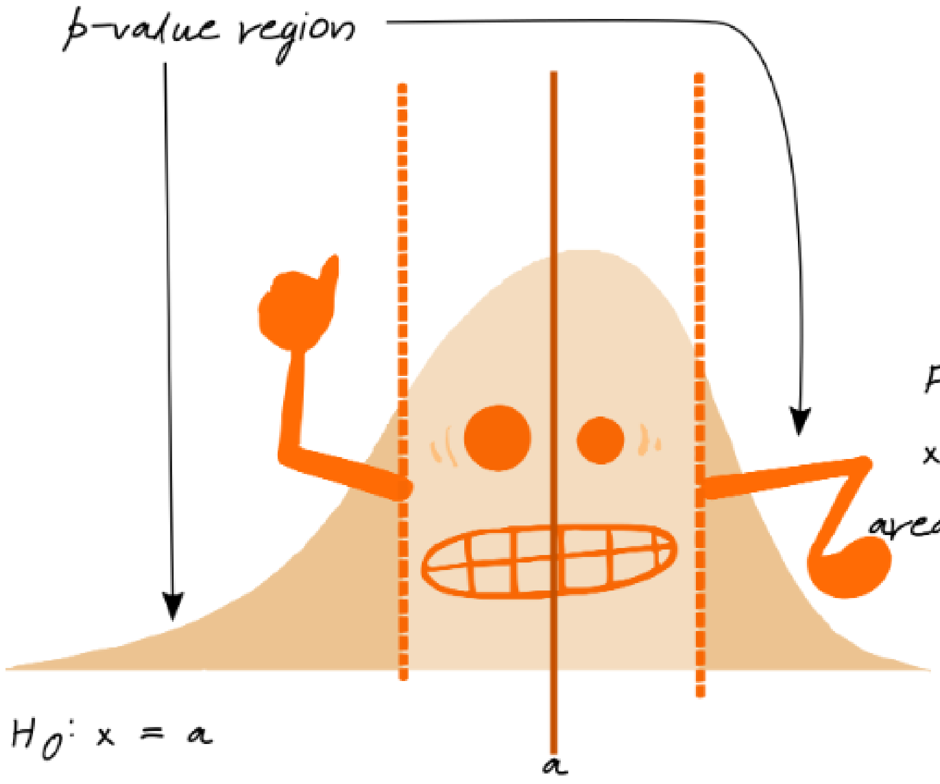
If the data from your sample seems unusual compared to this, then we have evidence!

- p-value $\leq \alpha \implies$ reject the null hypothesis and claim **we have** evidence in favor of the alternative hypothesis
- p-value $> \alpha \implies$ fail to reject the null hypothesis and claim **we do not have** evidence in favor of the alternative hypothesis.

Note: We NEVER accept the null hypothesis

P-VALUES

TWO-SIDED P-VALUES



H_0: x = a

H_A: x not equal to a

These tests are called two-sided tests

Because we don't care whether x < a or x > a - we only care if x = a.

For these tests, we have two regions where x is definitely not near a, so we count the area outside of both regions to get a p-value

If the p-value is less than 0.05 (or another pre-specified error level) then we reject H_0 and conclude that H_A is more likely.

Here, the p-value is higher than 0.05, so we don't get to reject H_0

CONFIDENCE INTERVALS

We have our **point estimate** from our observed summary statistic from our sample (this is our "best guess" for our population parameter).

Goal: Find a range of plausible values for our parameter (point estimate + uncertainty).

$\bar{x}$
Uncertainty

Confidence intervals are a range of values around the central estimate obtained from the sample data

CONFIDENCE INTERVALS

How do we find them? Use the quantiles (could call these percentiles) on from the **sampling distribution** either from bootstrapping or theory.

$$\text{point estimate} \pm \text{multiplier} \times \text{SE}$$

$$\bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

- If the interval falls completely above or below the **null value** $\implies$ reject the null and conclude evidence in favor of the alternative!

- If the interval falls contains the **null value** $\implies$ fail to reject the null and conclude we do not have evidence in favor of the alternative!

WHAT'S NEW?

Last week

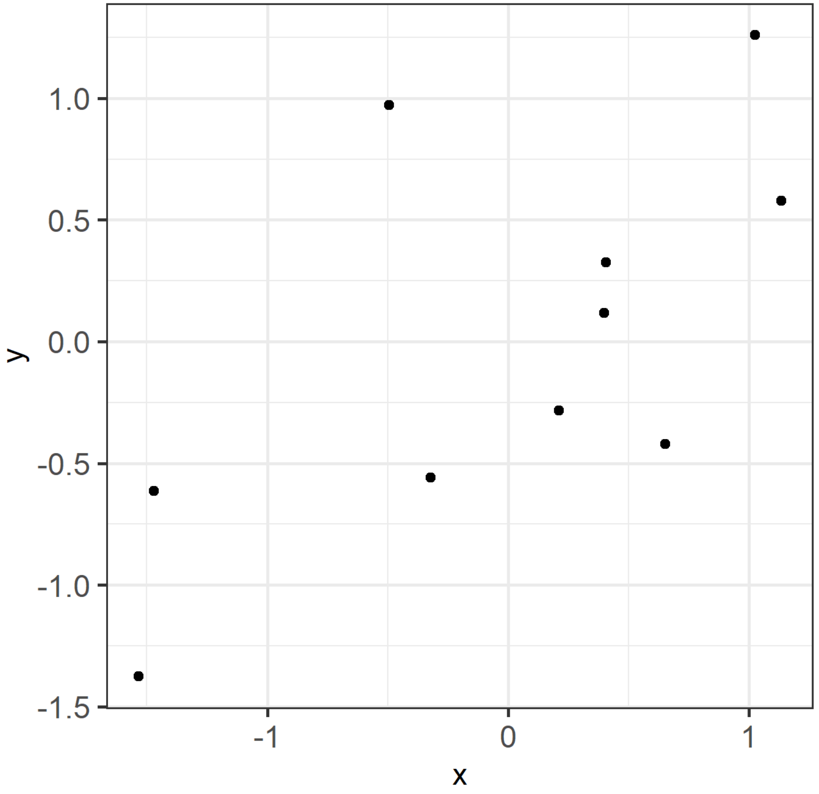
- ONE quantitative/numerical variable
- Summarize with measures of
    - center
        - mean/average, median/middle
    - spread
        - standard deviation - sd, inner quartile range - IQR = Q3 - Q1
- Visualize with: Dot-plots, histograms, box-plots

This week

- TWO quantiative/numerical variables
- Summarize with correlation (r) - measures strength of relationship/association
- Visualize with scatterplots

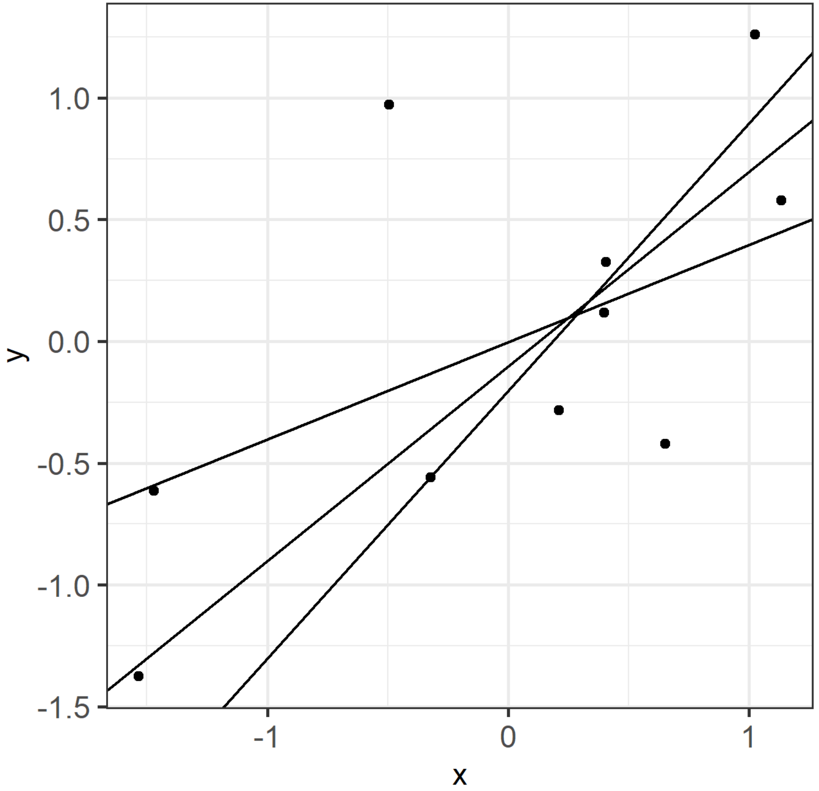
SIMPLE LINEAR REGRESSION

The principle of simple linear regression is to find the line (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.

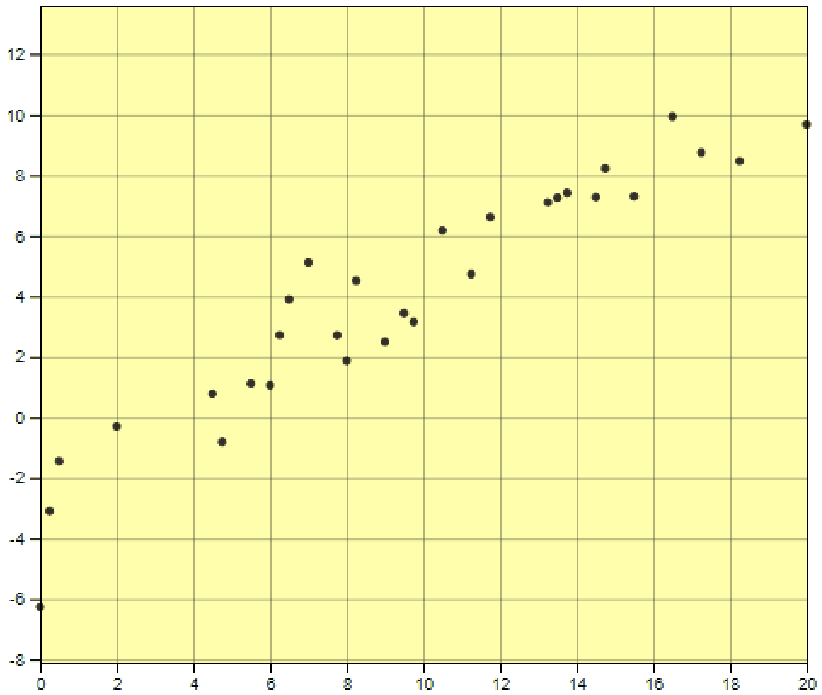


SIMPLE LINEAR REGRESSION

The principle of simple linear regression is to **find the line** (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.



Let's see your drawing skills!

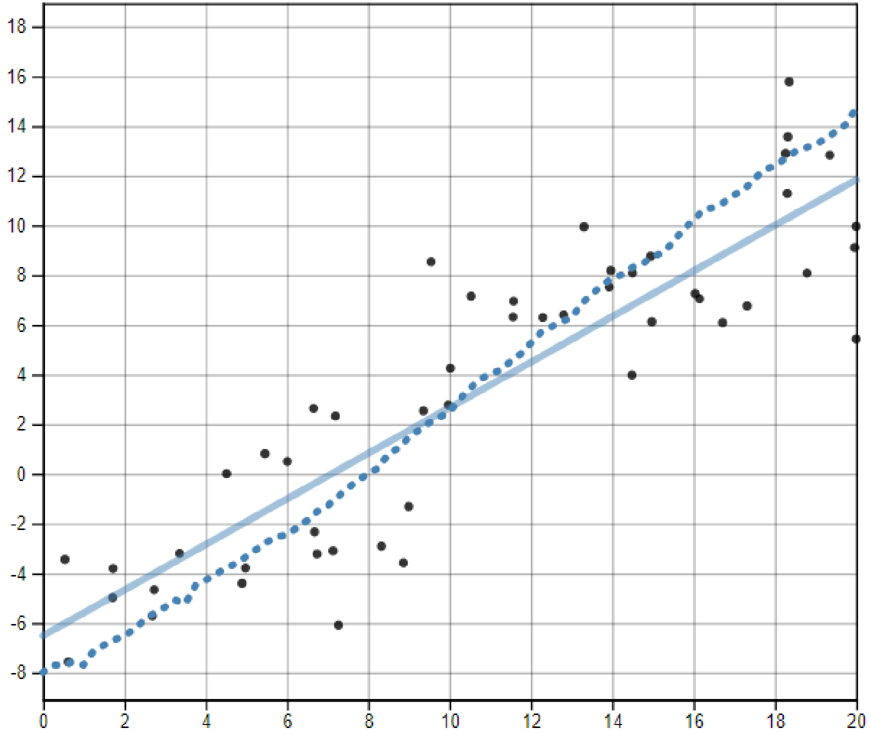


SCAN ME

OR VISIT bit.ly/3BF56Zj

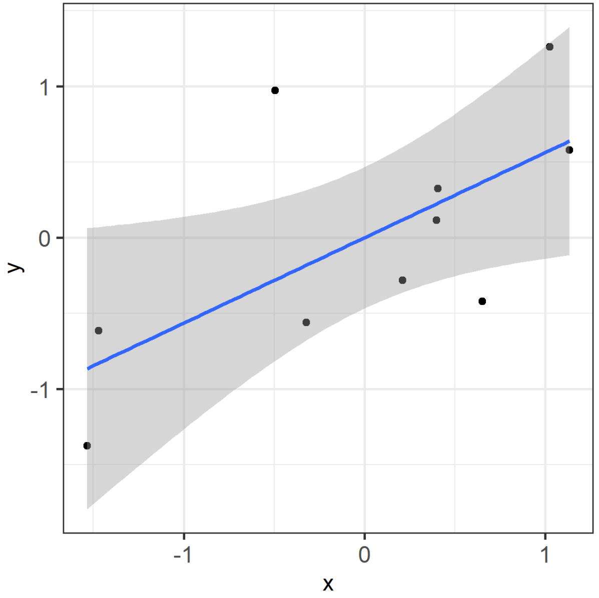
Did it looks something like this?

Did your neighbors look the same?



SIMPLE LINEAR REGRESSION

The principle of simple linear regression is to **find the line** (i.e., determine its equation) **which passes as close as possible to the observations**, that is, the set of points.



Equation of a line (y = mx + b):

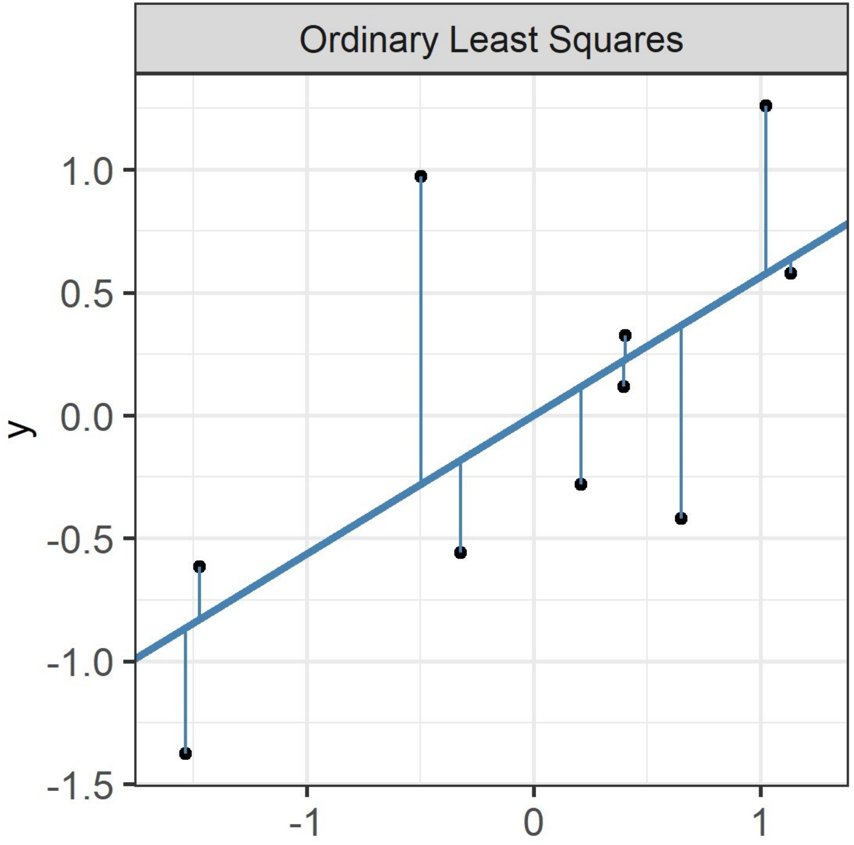
- Estimated from sample: $\hat{y} = b_0 + b_1 \times x$
- Population true line: $y = \beta_0 + \beta_1 \times x + error$

Recall how $\bar{x}$ was our point estimate for μ.

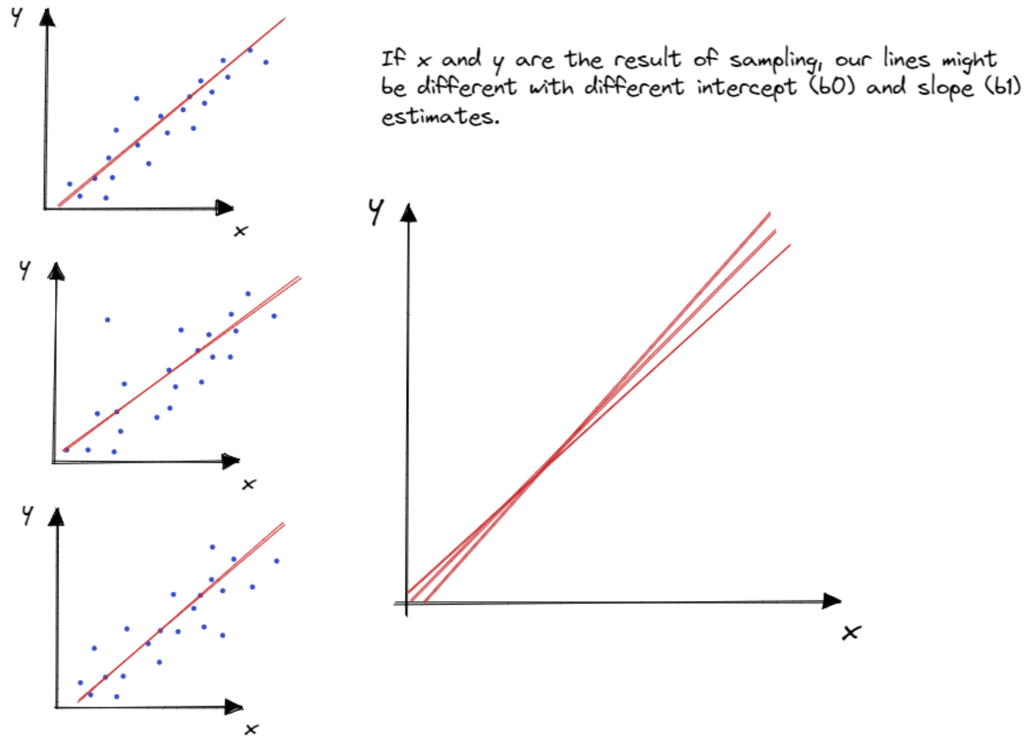
Big Idea: We use least squares regression (aka math) to find the "best" line.

RESIDUALS

Difference between observed points and predicted value (line) for the same x value.



INFERENCE FOR SLOPE



If x and y are the result of sampling, our lines might be different with different intercept (b0) and slope (b1) estimates.

Goals:

1. Test whether there is an association between x and y
2. Give a confidence interval for the true population slope

INFERENCE FOR SLOPE

Null: The true slope between x and y for the population is equal to 0 $(H_O : \beta_1 = 0)$

Alternative: The true slope between x and y for the population is different than 0 $(H_A : \beta_1 \neq 0)$

Then we look for evidence (p-values & confidence intervals)!

- Simulation - Randomization to create our null distribution
- Theory

DIVING PENGUINS

Diving Penguins

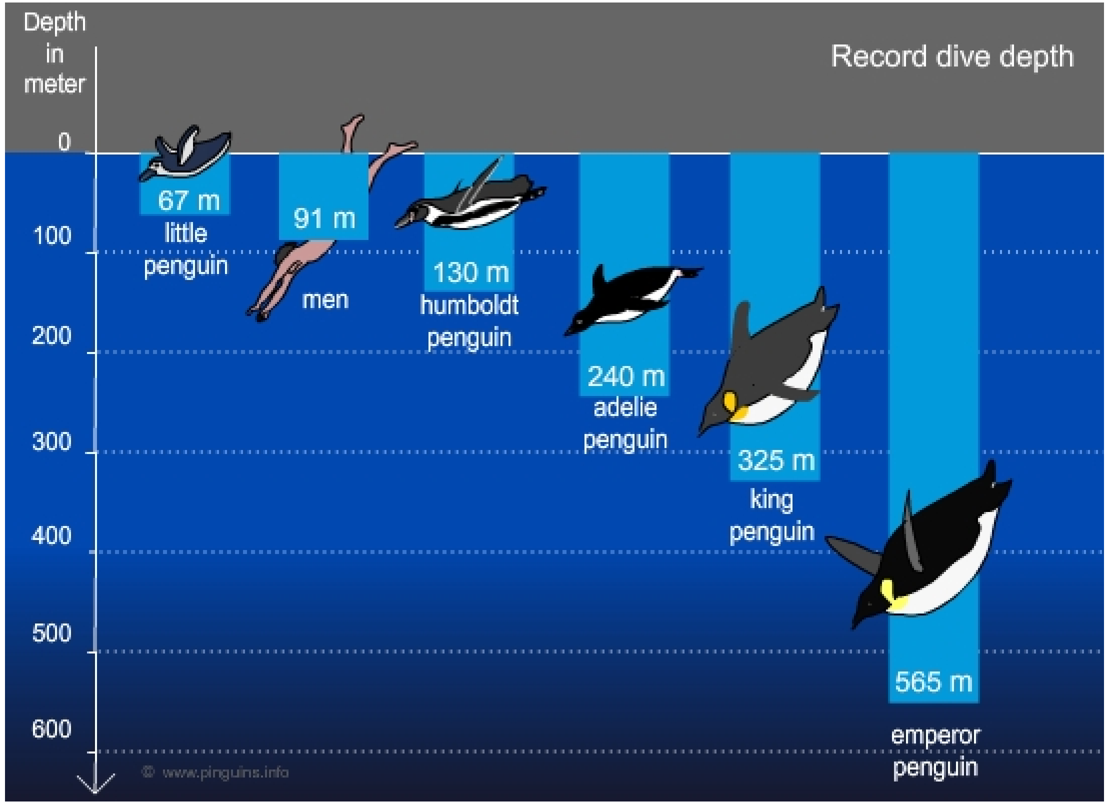


Image from: https://www.pinguins.info/Engels/Voortdiepte_eng.html