Intro to Statistics

Statistics with R

Basel R Bootcamp

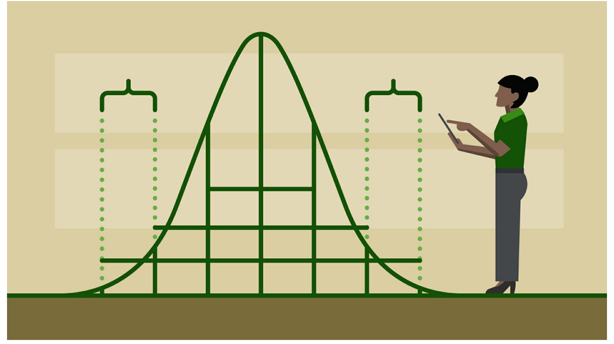
April 2019

Our goal in the next hour

In this hour, we will try to cover some of the **basic principles of statistical inference**.

1.
2.
3.
4.
5.
6.
7.
8.

This is a lot to cover, and it may not be clear from the beginning.
To help us, **we'll think about it in terms of beer**.



from cdn.lynda.com

from marketingweek.imgix.net

Example

Basel has many nice "Buvette's" that serve drinks in warm months.

The Oetlinger Buvette, one of our favorites, offers 33cl beers (or so they say...).

> I am convinced that the Oetlinger Buvette 'underpouring' its beers and they are not truly 33cl.

How can I find out?

How can I formulate my belief into an **formal hypothesis?**

How can I collect **data to test the hypothesis?**

What do you think?

from basel.com

from newsnetz.ch

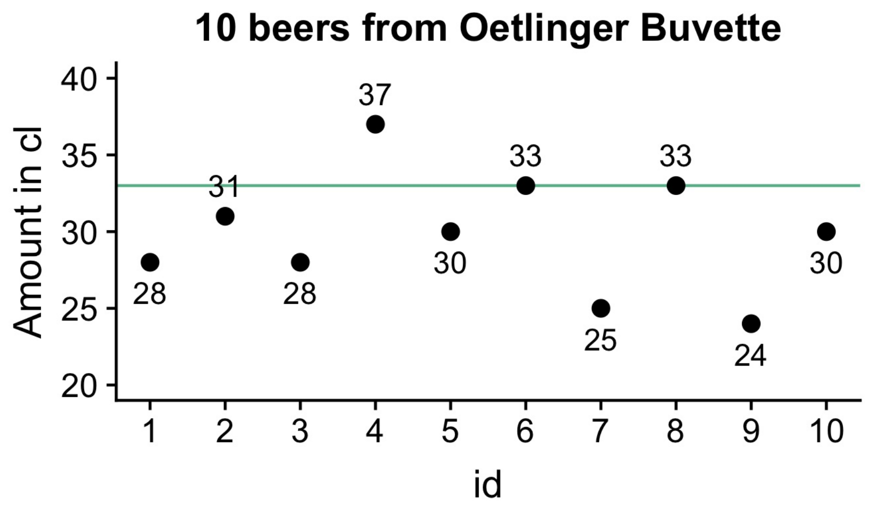
Beer hypothesis

The mean amount poured in 33cl beers by the Oetlinger Buvette is less than 33cl:

$$\Large H1: \mu < 33$$

Beer Data

I ordered 10 beers, and measured the exact amount in each cup, here are the results:



from basel.com

from newsnetz.ch

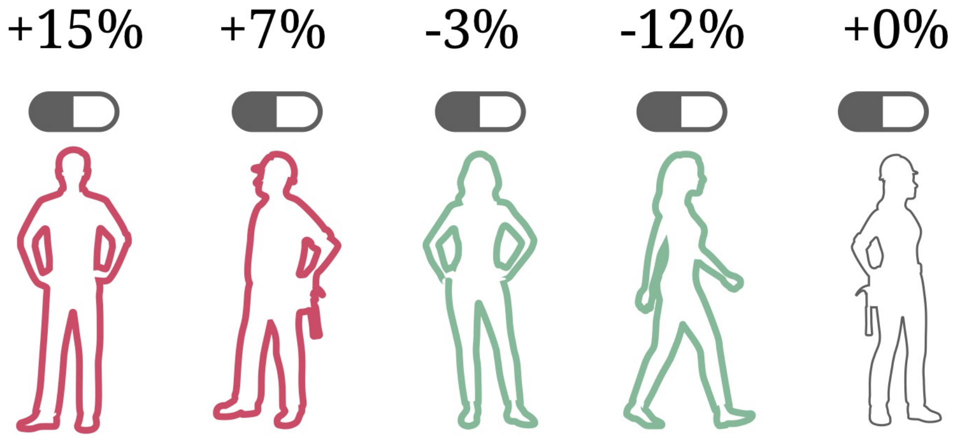
1. Variability

All interesting data processes have **variability**

- Stock prices change over time,
- Individual patients respond to drugs differently

Statistical inference is all about **accounting for variability**

If there was no variability, there would be no need to do statistics.

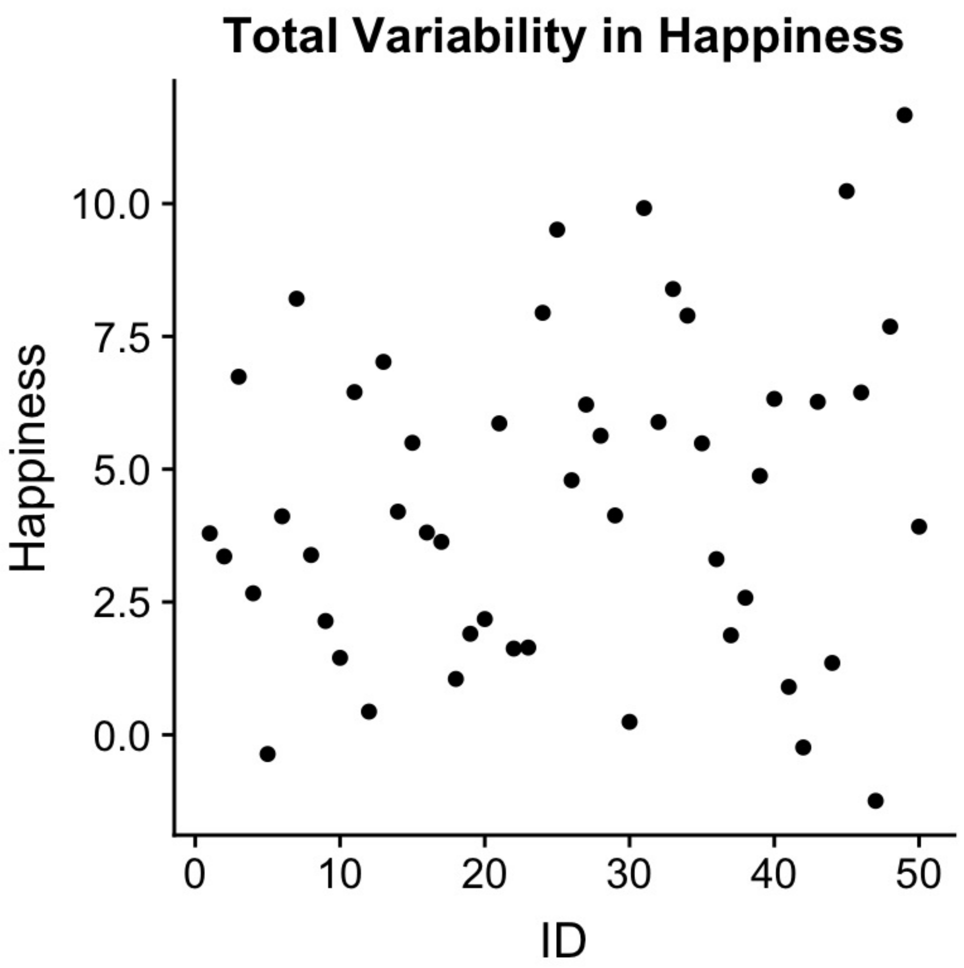


1. Variability

There are two types of variability: **systematic and unsystematic variability**.

Statistical inference typically seeks to **separate total variability into systematic and unsystematic portions**.

Variability Type	Definition
Systematic	Variation that **can** be explained by known variables
Unsystematic	Variation that **cannot** be explained by known variables

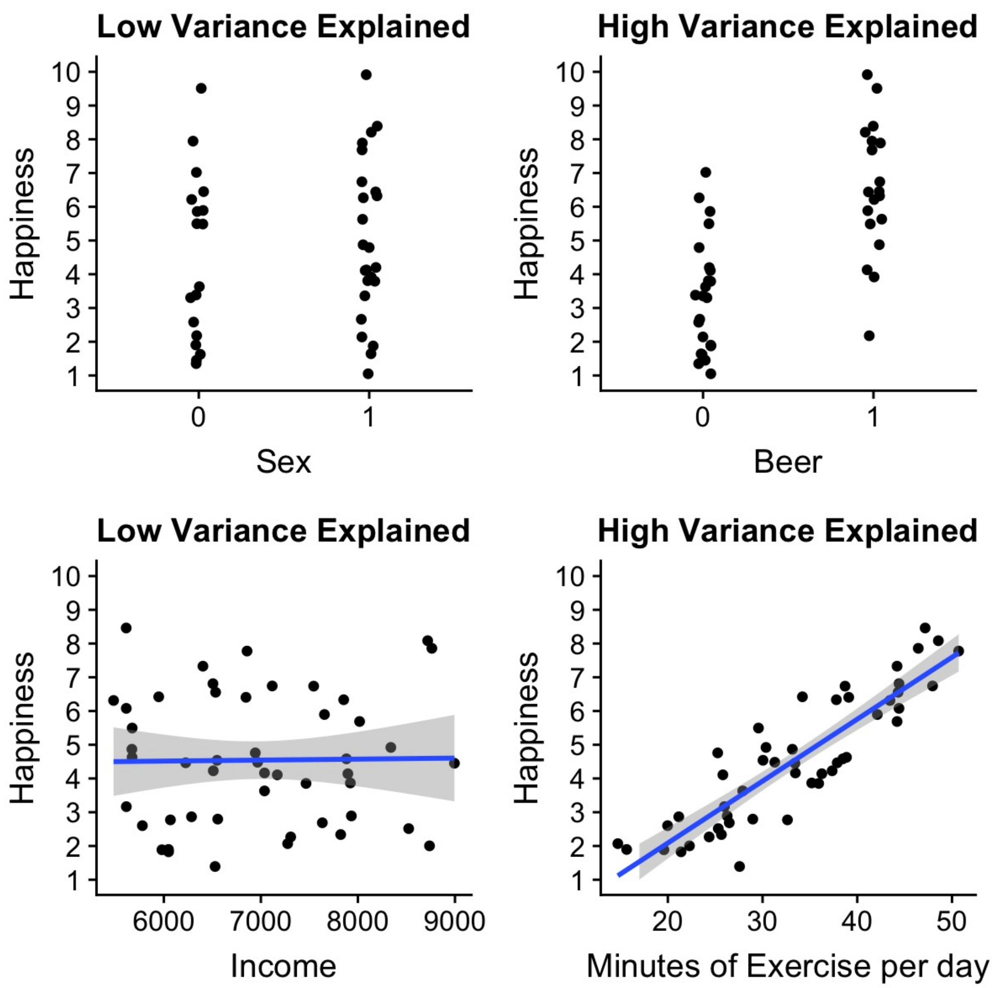


1. Variability

There are two types of variability: **systematic and unsystematic variability**.

Statistical inference typically seeks to **separate total variability into systematic and unsystematic portions**.

Variability Type	Definition
Systematic	Variation that **can** be explained by known variables
Unsystematic	Variation that **cannot** be explained by known variables

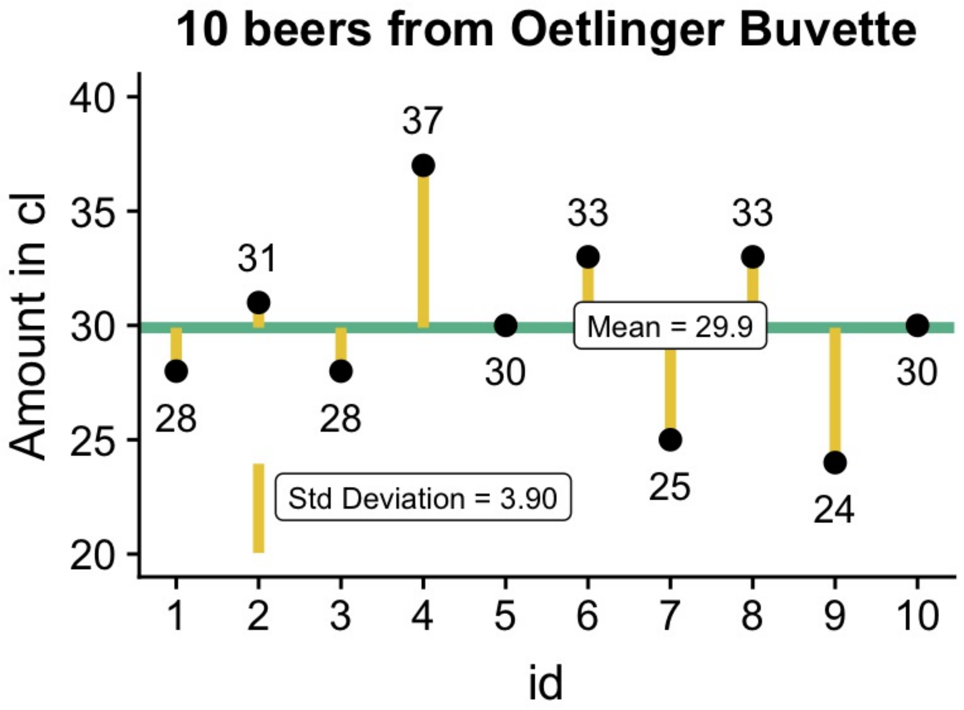


2. Sample Statistics

Once we collect data, we always look for ways to **summarise** the data into **sample statistics**

Sample statistics give us **estimates** of key model parameters (more on this later) and usually (but not always) fall into one of two types:

Type	Examples
Central Tendency	Mean, mode, median
Variability	Standard deviation, variance, range



$$Mean = \frac{28+31+28+...}{10} = 29.9$$

$$Stand. \; Dev. = \sqrt{\frac{(28-29.9)^2+(31-29.9)^2+...}{10-1}} = 3.90$$

3. Sampling Procedures

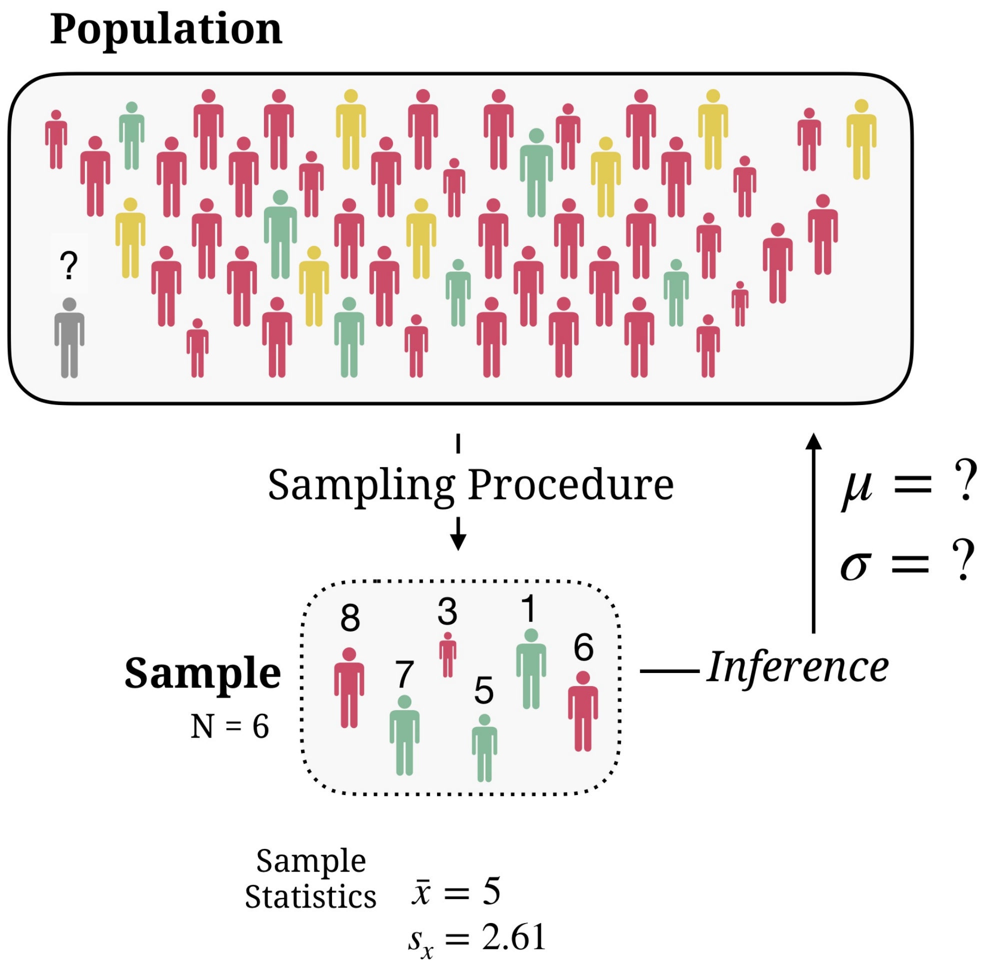
In statistics, we always distinguish between a **sample** and a population

Populations are what we are really interested in

- How will be affected by Drug X?
- How will change?
- Or is the pouring amount different from 33cl?

To learn about the population, we must use **sampling procedures** to obtain a **sample** of cases.

Then, inferential statistics allows us to **inferences** to populations.



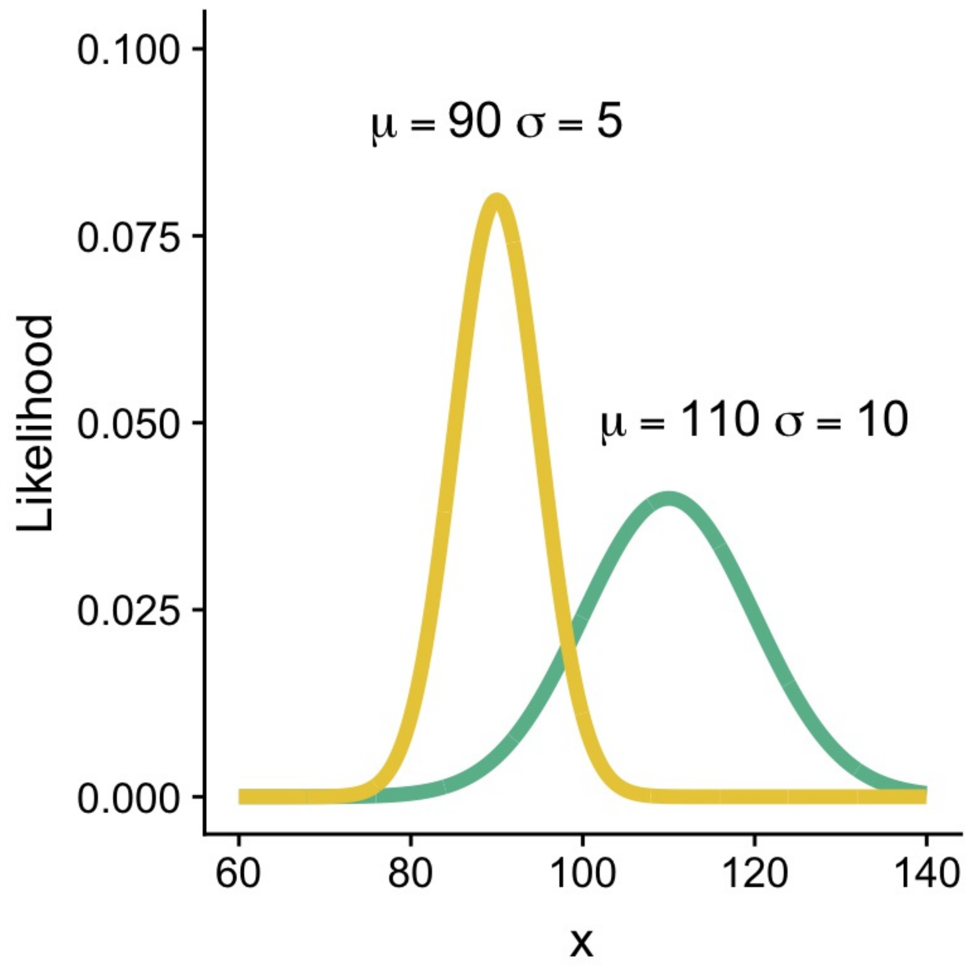
4. Distributions

(Parametric) statistics is built on **calculating the likelihood of data** given a probability distribution.

A probability distribution is a **mathematical formula** that precisely defines **how likely** every possible value in a dataset is. **List of distributions**.

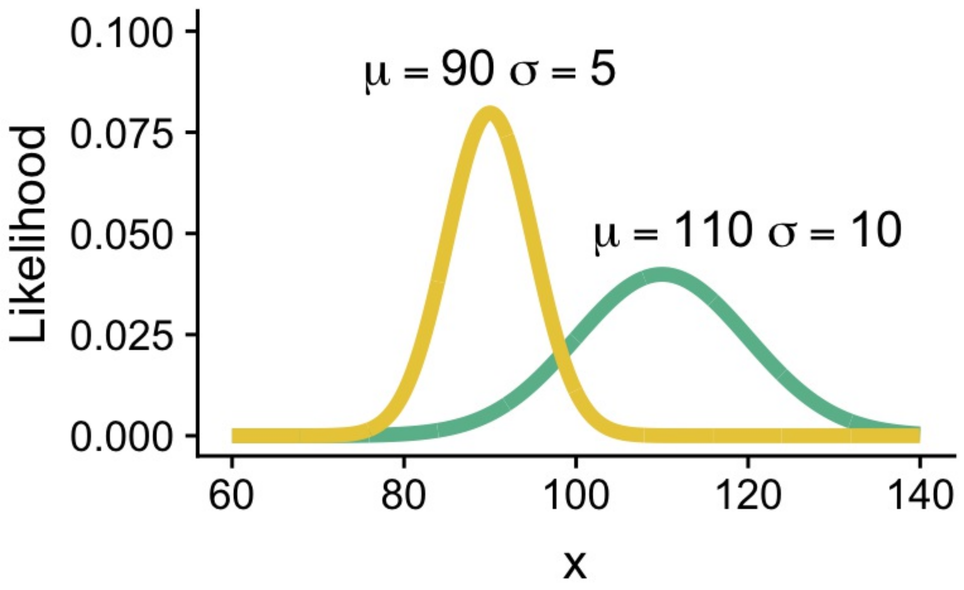
3 key aspects of a distribution

1. **Probability Density Function (PDF)** - Formula defining the distribution (R knows these)
2. **Support** - What values can x take on?
3. **Parameters** - Values that allow you to change the shape of the distribution? (e.g.; mean and variability?)



Normal Distribution

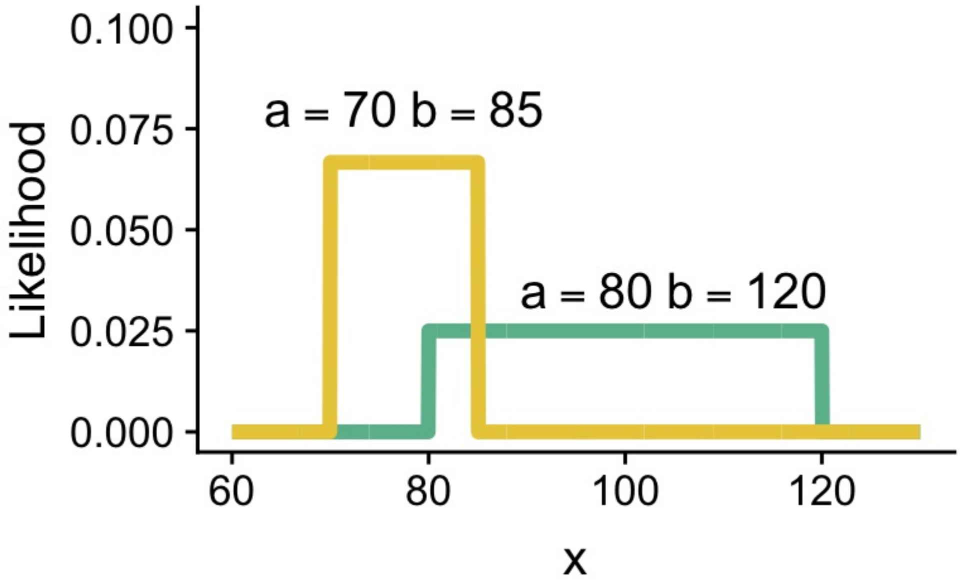
A aka Gaussian distribution, which is the **most important distribution in all of statistics**.



Aspect	Formula
	$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\mu)^2/2\sigma^2}$$
	$$x \ \epsilon(-\infty, \infty)$$
	μ (Center; mean)
	σ (Variability; stand. dev.)

Uniform Distribution

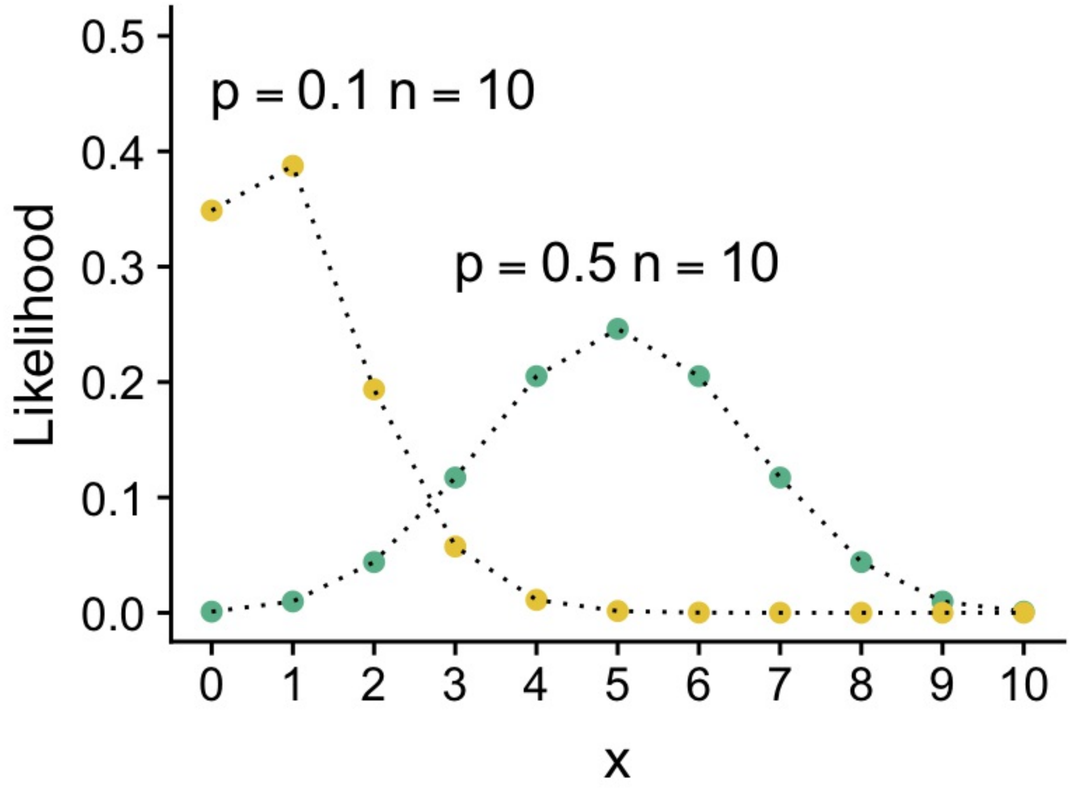
A 'Flat distribution', used **when everything is equally likely**, within a range.



Aspect	Formula
	$f(x) = \frac{1}{b-a}$
	$x \ \epsilon (a, b)$
	a (Minimum)
	b (Maximum)

Binomial Distribution

A discrete "Counting" distribution answering: If I flip a coin N times, with p(Head) = p, how many times will I get heads?



Aspect	Formula
	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$
	$x \; \epsilon \{0, 1, ...n\}$
	p (p(success))
	n (No. trials)

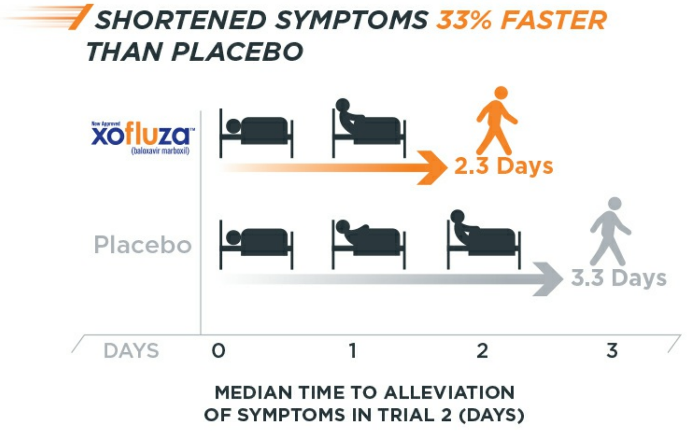
5: Likelihood

Why do we need distributions? To calculate **likelihoods** of data.

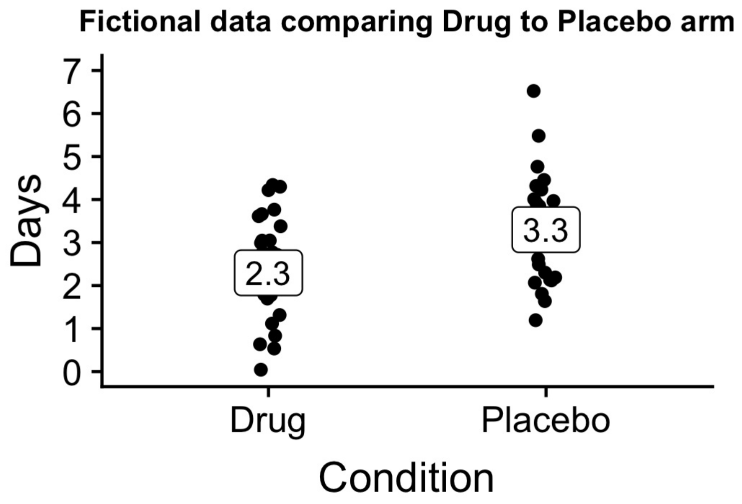
> How likely is it that I would get this trial result if the drug is   better than a placebo?

Knowing this likelihood allows us to **fit parameters**, **test** models, and make **predictions** about future data

> Given that out of 50 trial patients, the average recovery time was 2.3 days, what is the most likely distribution of recovery times for future patients?



An ad for xofluza, from xofluza.com



5: Likelihood

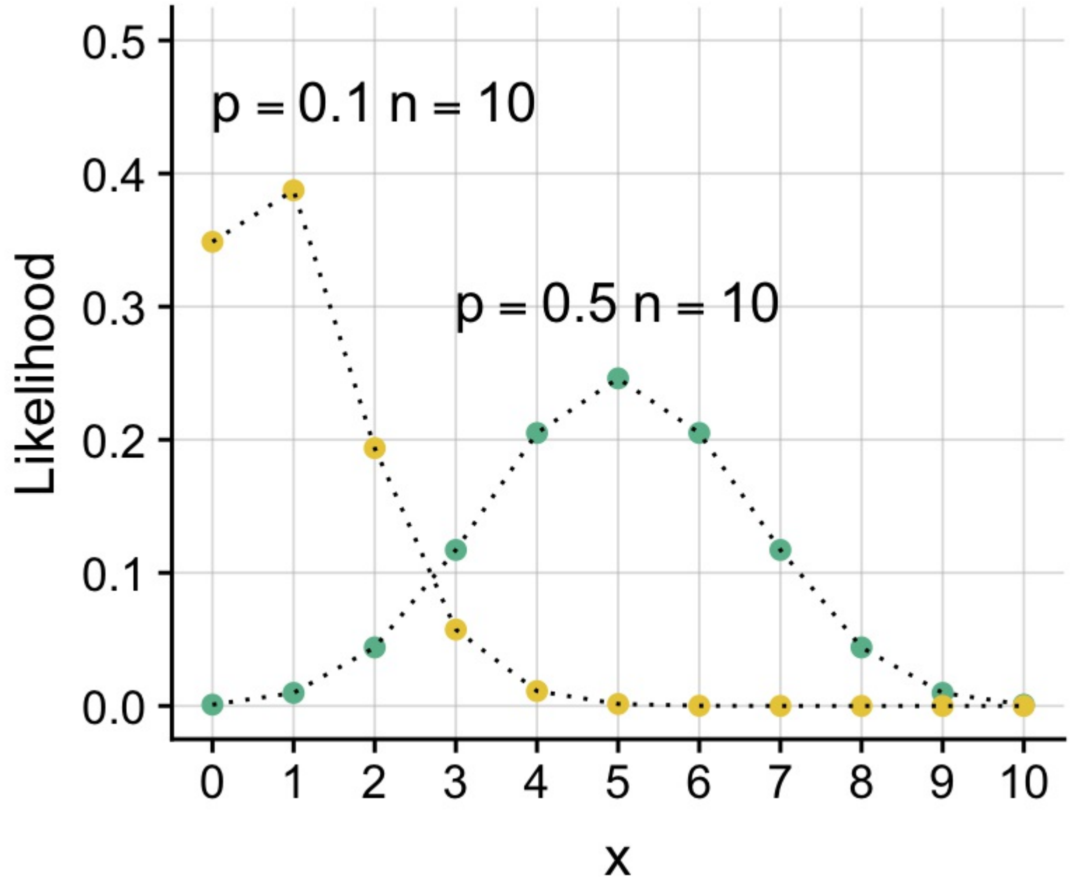
Using the binomial distributions on the right, answer the following questions:

Q1

If there is a 50% chance of a clinical trial being successful, then out of 10 drugs, **how likely is it that exactly 5 will be successful**?

Q2

If there is a 10% chance that a customer will default on his/her loan, then out of 10 customers, **how likely is it that none (0) will default**?

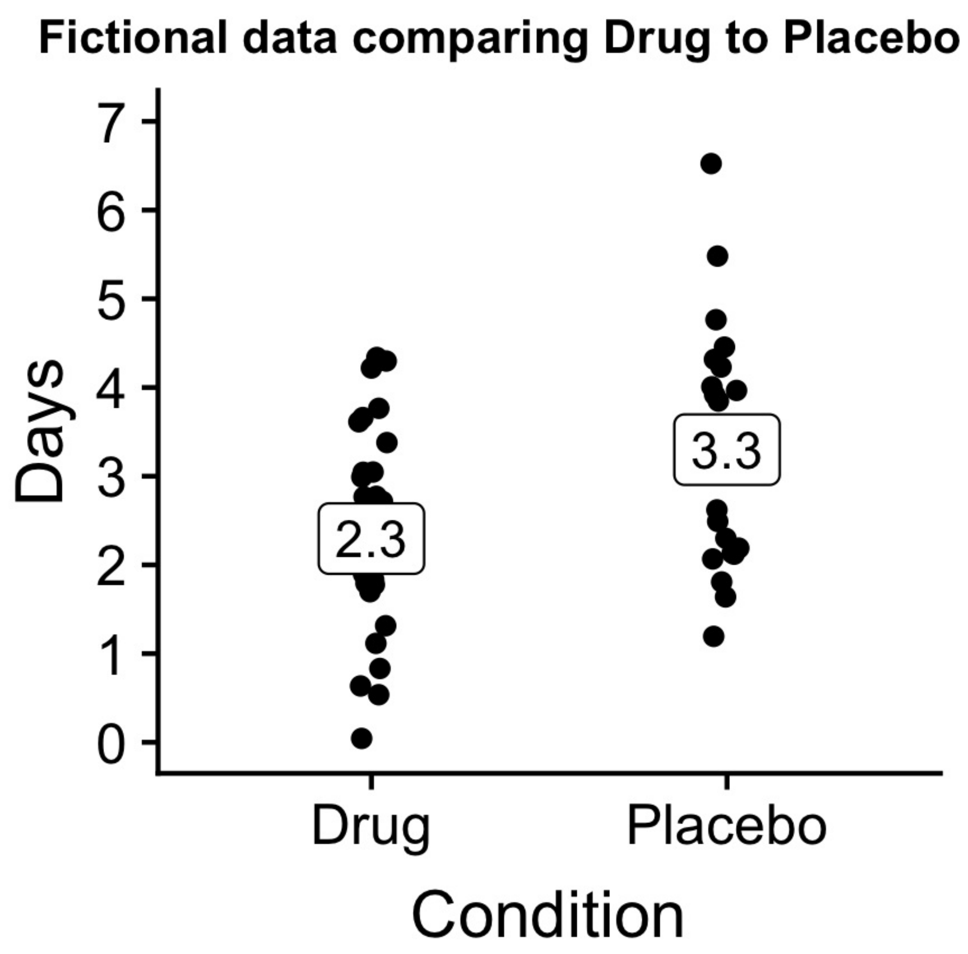


6: Null hypothesis testing

Null hypothesis testing is a statistical framework where one hypothesis (H_0) is tested to defend the other, alternative hypothesis (H_1).

This evaluation is performed by calculating the likelihood of obtaining the data **assuming** that the null hypothesis true.

Hypothesis	Description	Example
Null (H_0)	A proposed effect **does not exist** and variation **is not systematic**.	Drug and placebo have the same effect.
Alternative (H_1)	A proposed effect **does exist** and variation **is systematic**	Drug and placebo do *not* have the same effect

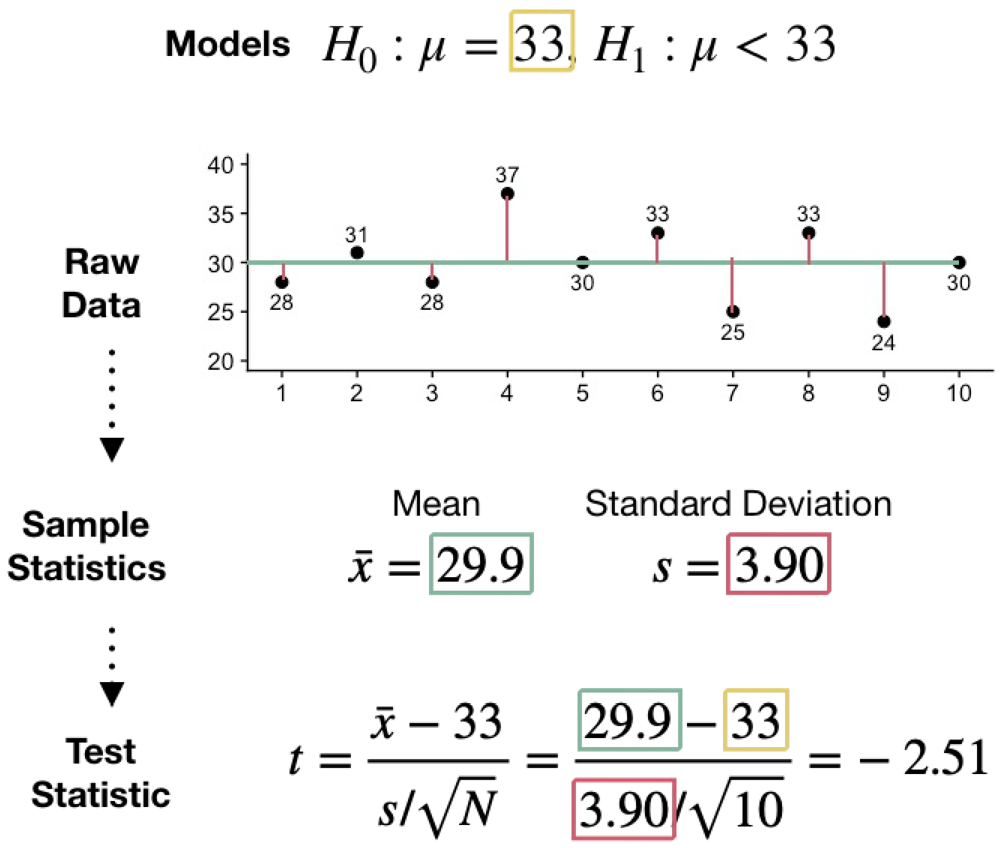


7: Test statistics

Sample statistics (like means and standard deviations) are converted into **test statistics**.

Test statistics are unit-free numbers that help you quantify how likely data is given a null hypothesis. The **more extreme** (i.e.; highly positive or highly negative) your test statistic is, the **more evidence against** the null hypothesis.

Test	Test statistic
t-test	t-statistic
Correlation test	Correlation coefficient
Binomial	Number of successes



8: P-value

-values are used to quantify the likelihood of data given the **probability distribution** under the **null hypothesis (H0)**.

If a p-value is (i.e.; $p < .05$), this means that the likelihood of obtaining that data given the null hypothesis is , suggesting that the null hypothesis is .

Formally

A p-value is the probability of obtaining a test statistic as extreme or more extreme than what you got assuming a null hypothesis is true

Decision rule

$p < .05 \rightarrow$ Reject H_0

Test Statistic		p-value	Likelihood if H0 is true
0		0.95	Extremely
.		0.90	
.		0.85	
.		0.80	
.		0.75	
Small		0.70	Very
.		0.65	
.		0.60	
.		0.55	
Medium	→	0.50	Coin-toss
.		0.45	
.		0.40	
.		0.35	
.		0.30	
Large		0.25	Unlikely
.		0.20	
.		0.15	
.		0.10	
Huge		**0.05**	**Very unlikely**
Inf		**0.00**	

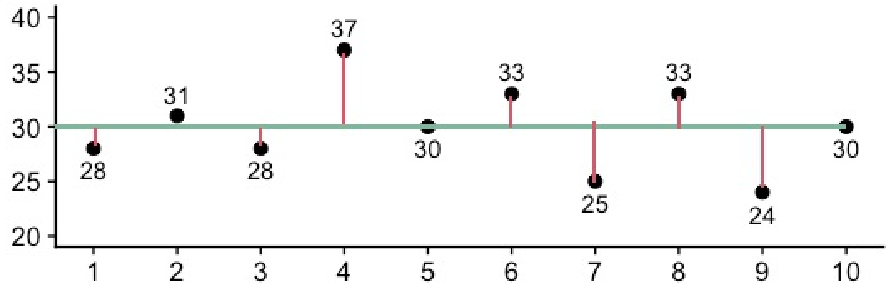
What about Oetlinger?

Step	Result
H_0	The mean amount of beer poured by Oetlinger is 33ml.
Sample statistics	Mean = 33 Std. Deviation = 3.90.
Test and - value	We calculated, based on the sample statistics, a test statistic of **=-2.51** and a **p-value** of

Conclusion - Using a $p < 0.05$ threshold, we conclude that the null hypothesis is likely wrong and that…the Oetlinger buvette is pouring less than 33cl!

Models $H_0 : \mu = \boxed{33},\ H_1 : \mu < 33$

Raw Data



Sample Statistics

Mean $\bar{x} = \boxed{29.9}$ Standard Deviation $s = \boxed{3.90}$

Test Statistic

$$t = \frac{\bar{x} - 33}{s/\sqrt{N}} = \frac{\boxed{29.9} - \boxed{33}}{\boxed{3.90}/\sqrt{10}} = -2.51$$

P-value

$$p = 0.0273$$

What about Oetlinger?

Step	Result
H_0	The mean amount of beer poured by Oetlinger is 33ml.
Sample statistics	Mean = 29.9 Std. Deviation = 3.90.
Test and - value	We calculated, based on the sample statistics, a test statistic of =-2.51 and a **p-value** of

Conclusion - Using a $p < 0.05$ threshold, we conclude that the null hypothesis is likely wrong and that...the Oetlinger buvette is pouring less than 33cl!

T-test in R

```
# Define beer sample
beer <- c(28, 31, 28, 37, 30,
          33, 25, 33, 24, 30)

# Conduct one-sample t-test
t.test(x = beer,    # Sample values
       mu = 33)     # Null Hypothesis
```

```
##
##     One Sample t-test
##
## data:  beer
## t = -2.5, df = 9, p-value = 0.03
## alternative hypothesis: true mean is not equal to 33
## 95 percent confidence interval:
##  27.11 32.69
## sample estimates:
## mean of x
##      29.9
```

Questions?

Schedule