Statistical Testing Using R

Mark J.H. Ou

PhD Student

IMBIM, Uppsala University

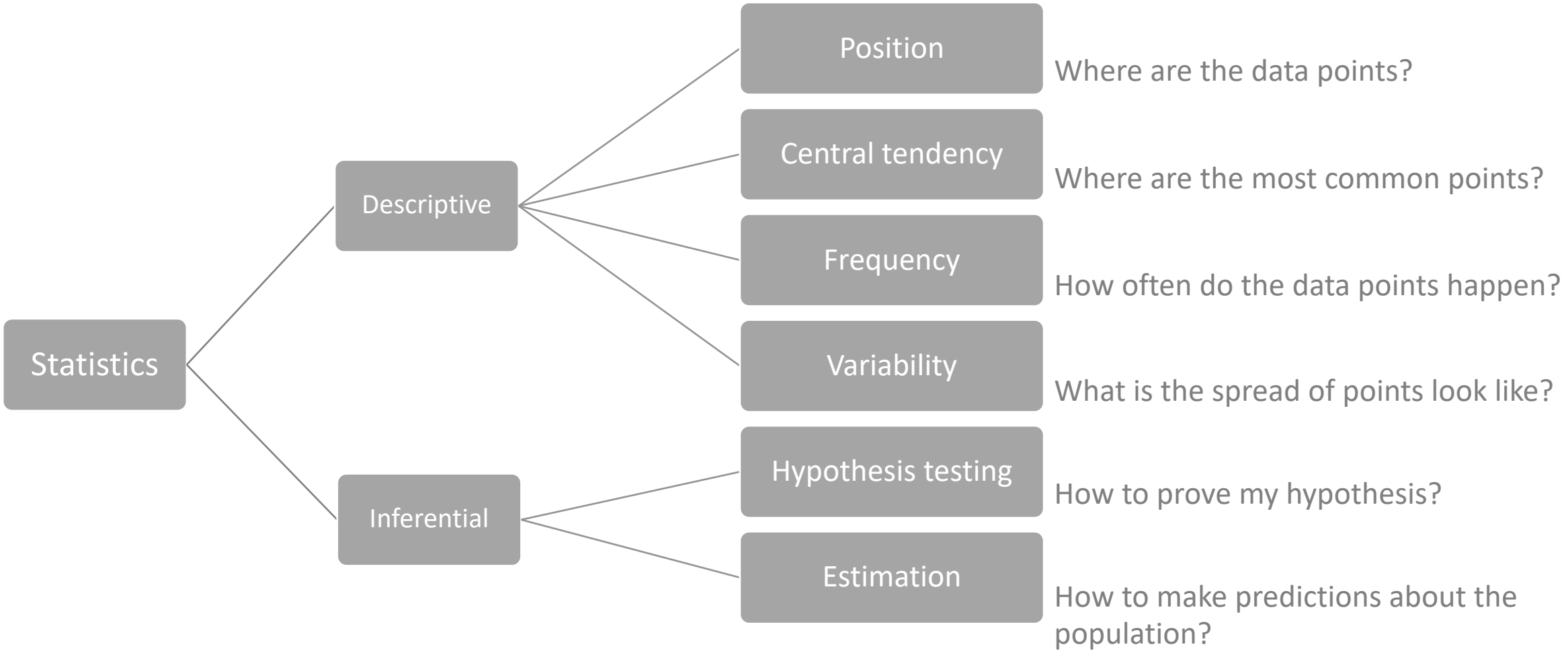
Why statistics

Sugar content in degree saccharine

- Too much ➔ Pay more tax
- Too little ➔ alcohol level too low
- Small sample size

$$Z = \frac{(M - \mu)}{\frac{\sigma}{\sqrt{N}}} \rightarrow \frac{(M - \mu)}{\frac{SD}{\sqrt{N}}}$$

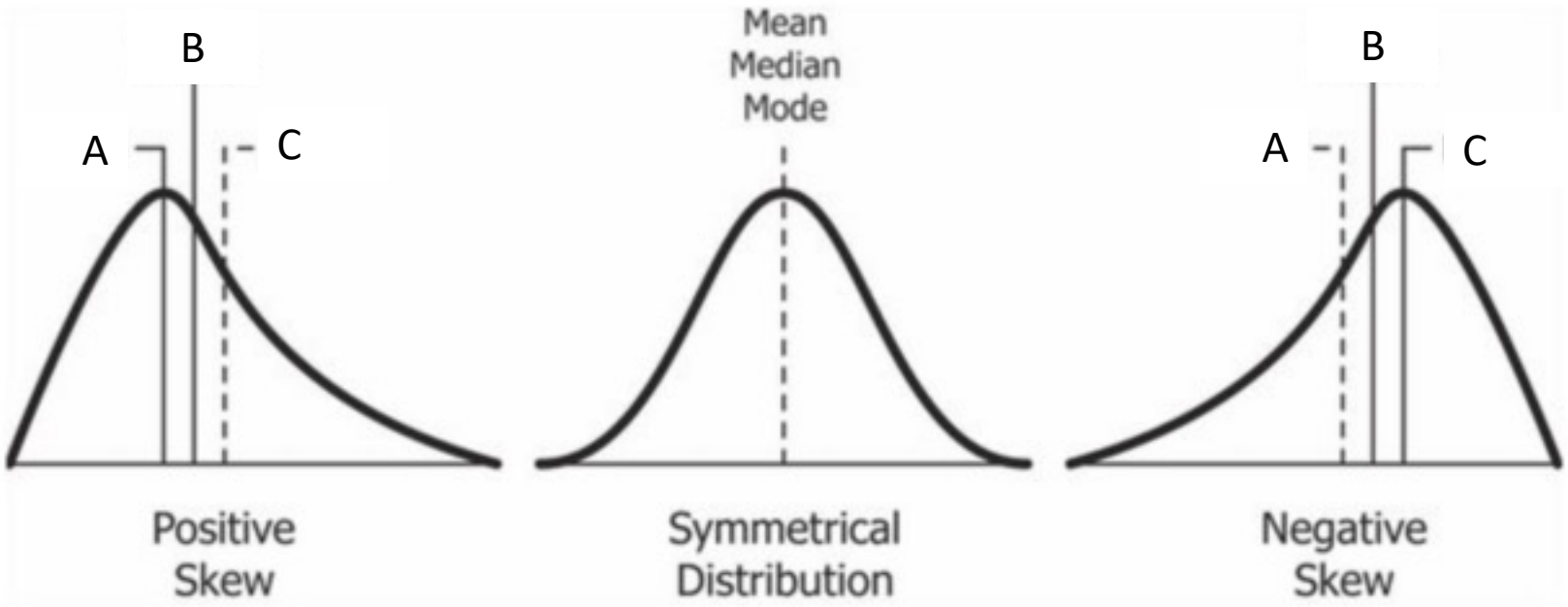
Type of Statistics



Descriptive statistics

Central tendency

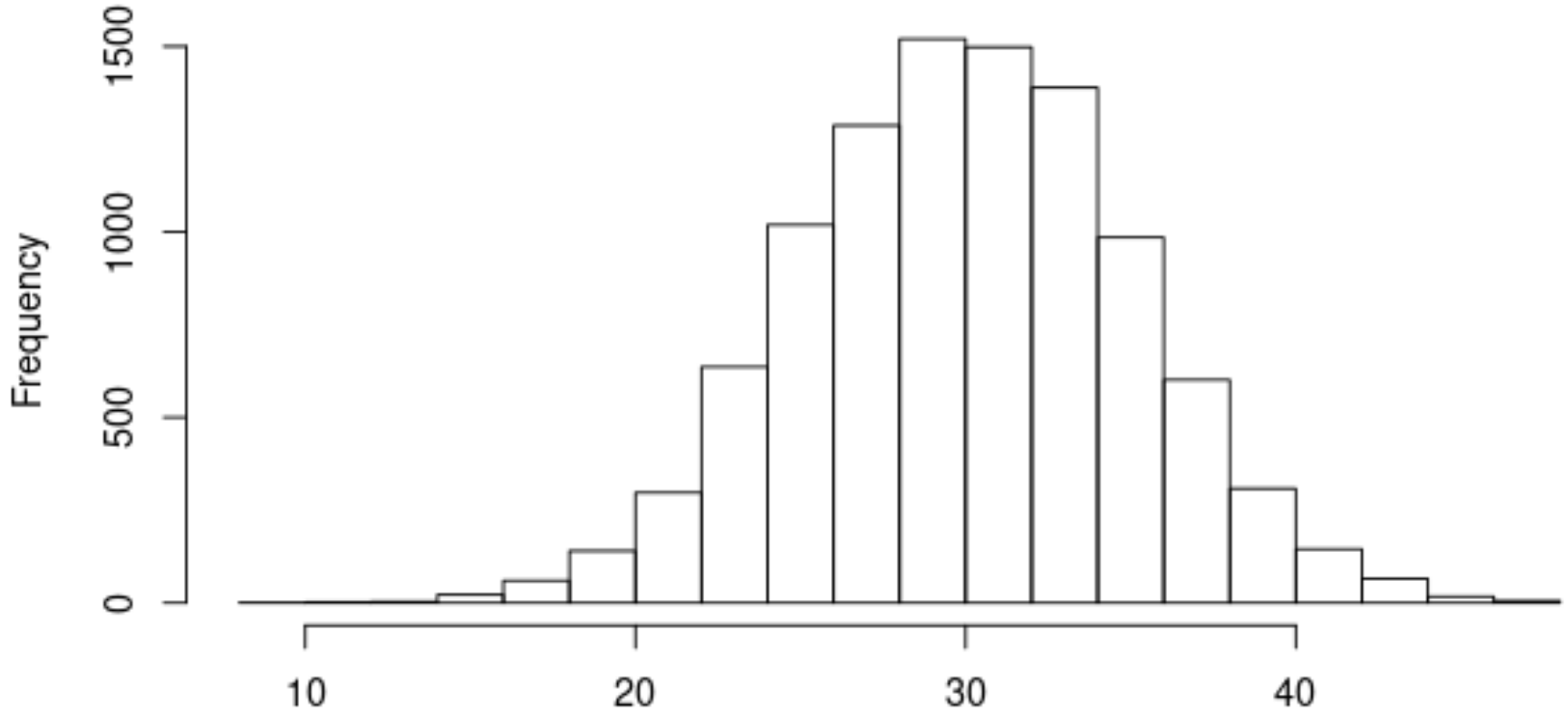
- Mean: the sum of all values divided by the total number of values
- Median: the middle number in an ordered data set
- Mode: the most frequent value



Central tendency in R

- Mean: `mean(x)`
- Median: `median(x)`
- Mode: no direct function to use, `table(x)`

Measurement of dispersion



Measurement of dispersion

- Range and Interquartile range (IQR)
  - IQR = $Q_3 - Q_1$
- Mean deviation

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

- Variance

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Measurement of dispersion

- Range and Interquartile range (IQR)
  - Range: `diff(range(x))`
  - IQR: `IQR(x)`
- Mean deviation

  `mad(x)`
- Variance

  `var(x)`
- Standard deviation

  `sd(x)`

Distributions

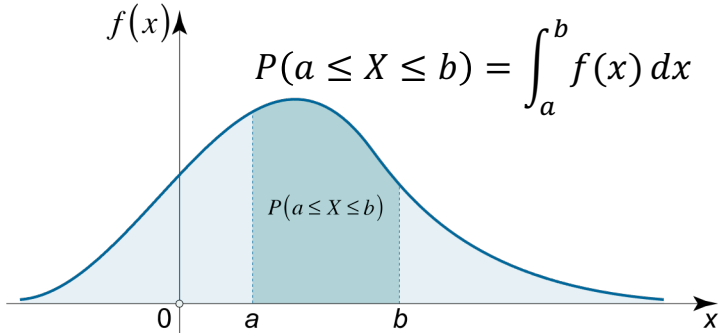
Continuous probability distribution

- Lots of biological measurements are continuous
  - Height
  - Body weight
  - Blood pressure

- Questions we wanted to answer are
  - Probability of an adult man whose body weight is not greater than 65 kg
    ➔ $P(X \leq 65)$
  - Probability of a person's body weight is between 55 and 65 kg
    ➔ $P(55 \leq X \leq 65)$

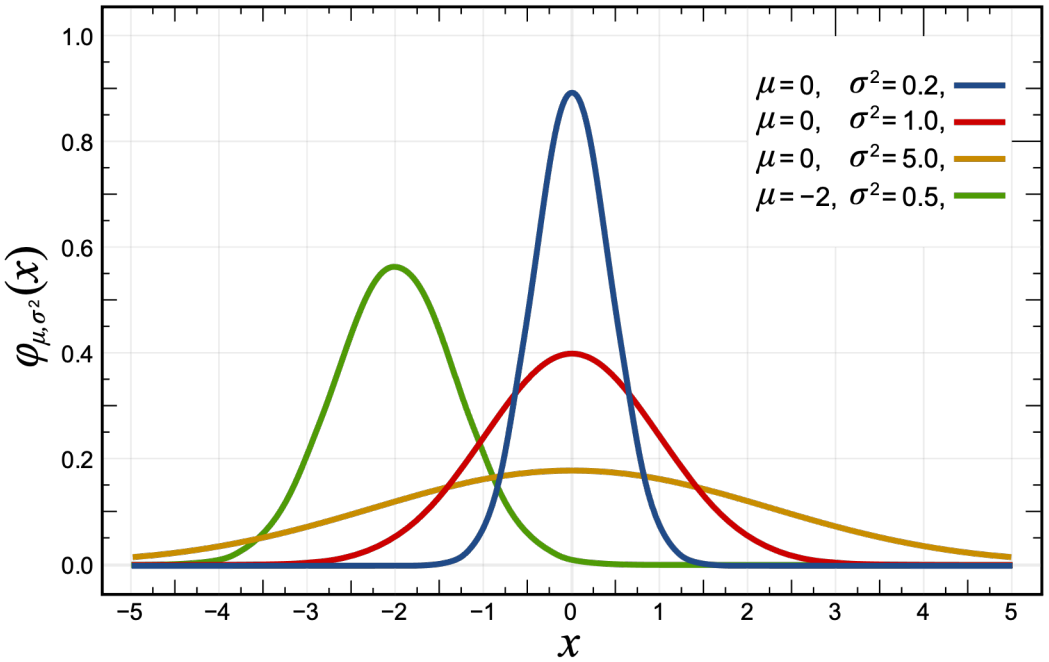
Probability density function (p.d.f.)

- Continuous random variable
- Value at any given sample in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would be close to that sample
- Every continuous random variable (X) has a p.d.f, written as $f(x)$, that satisfies the following conditions:
  - $f(x) \geq 0$, for all x ➔ non-negative everywhere
  - $\int f(x)dx = 1$ ➔ Integral over the entire space is equal to 1
  - What is the probability of $P(X = a)$?



Normal distribution

- Continuous distribution
- Widely used in the natural and social science



p.d.f. of Normal distribution

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

- μ: The mean of the distribution
- σ: The standard deviation of the distribution

Common notation

$$X \sim N(\mu, \sigma^2)$$

p.d.f. = probability density function

Standard normal distribution

- A normal distribution with a mean of 0 and a standard deviation of 1

$$X \sim N(0, 1)$$

- Standard score: Z score

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

Calculate Probability

- $X \sim N(60, 5^2)$
- Calculate $P(55 \leq X \leq 62) = ?$

By hand

$$
\begin{aligned}
&P(55 \leq X \leq 62) \\
&= P\left(\frac{55-60}{5} \leq \frac{x-\mu}{\sigma} \leq \frac{62-60}{5}\right) \\
&= P(-1 \leq Z \leq 0.4) \\
&= P(Z \leq 0.4) - P(Z \leq -1) \\
&= 0.6554 - 0.1587 \\
&= 0.4967
\end{aligned}
$$

Normal table available HERE

In R

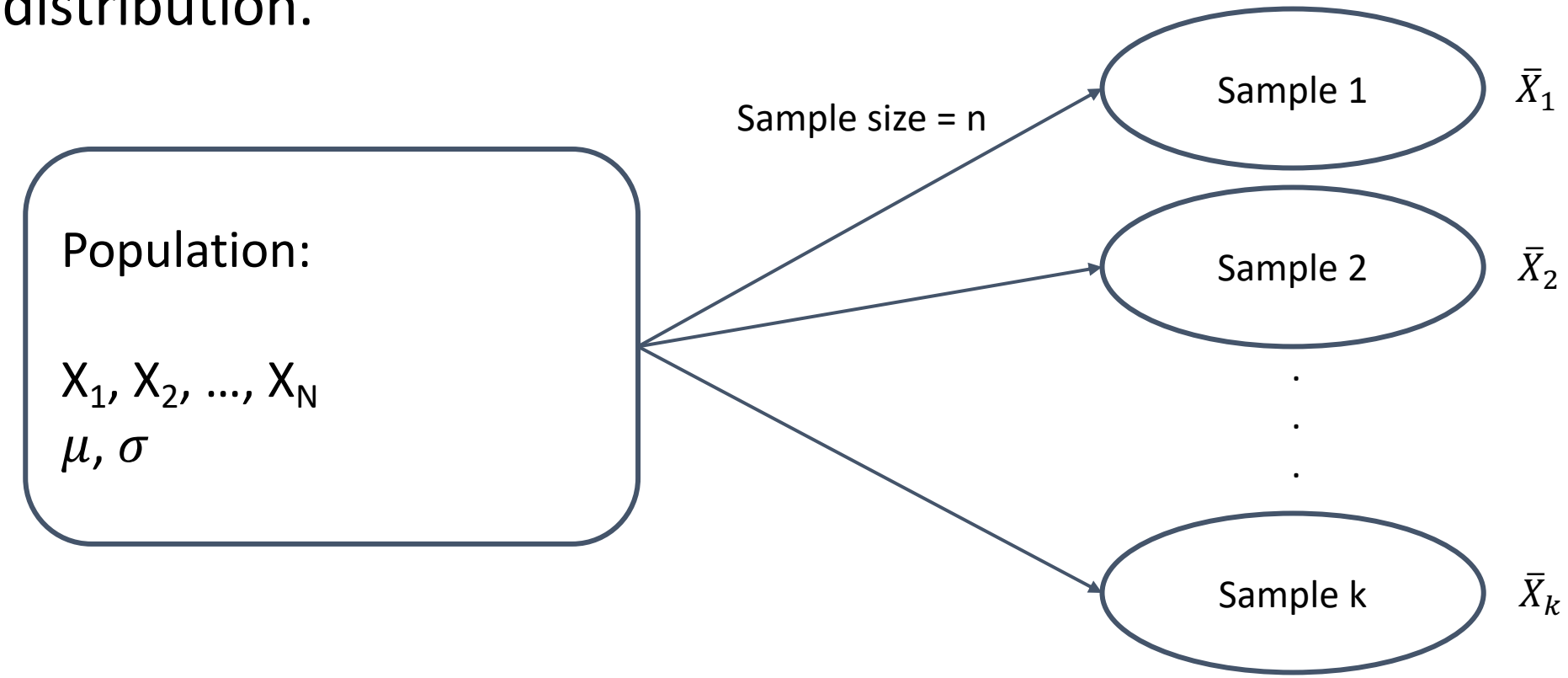
```
pnorm(q, mean, sd, lower.tail=TRUE)
```

- q: vector of quantiles
- mean: Population mean
- sd: Population SD

pnorm(62, 60, 5) - pnorm(55, 60, 5)

Samling Distribution

Sampling distributions are important for inferential statistics. We collect sample data and estimate parameters of the population distribution.



An example of sampling distribution

Population
[2, 4, 6]

Population

Observed	Probability
2	1/3
4	1/3
6	1/3

Sampling (n = 2) WITH REPLACEMENT

Observed	Probability
2	1/9
3	2/9
4	3/9
5	2/9
6	1/9

Population mean vs. Sample mean

Population mean

Observed x	Probability $P(x)$	$x \cdot P(x)$
2	1/3	2/3
4	1/3	4/3
6	1/3	6/3
		Mean = 12/3 = 4

Sample mean

Sample mean $\bar{x}$	Probability $P(\bar{x})$	$\bar{x} \cdot P(\bar{x})$
2	1/9	2/9
3	2/9	6/9
4	3/9	12/9
5	2/9	10/9
6	1/9	6/9
		Mean = 36/9 = 4

The population mean is equal to the sample mean ➔ Unbiasedness

Sampling distribution

- The distribution of all sample means
- Symmetrical distribution
- The mean of the sampling distribution is equal to the population mean
- The variance of sampling distribution vs. population is shown as

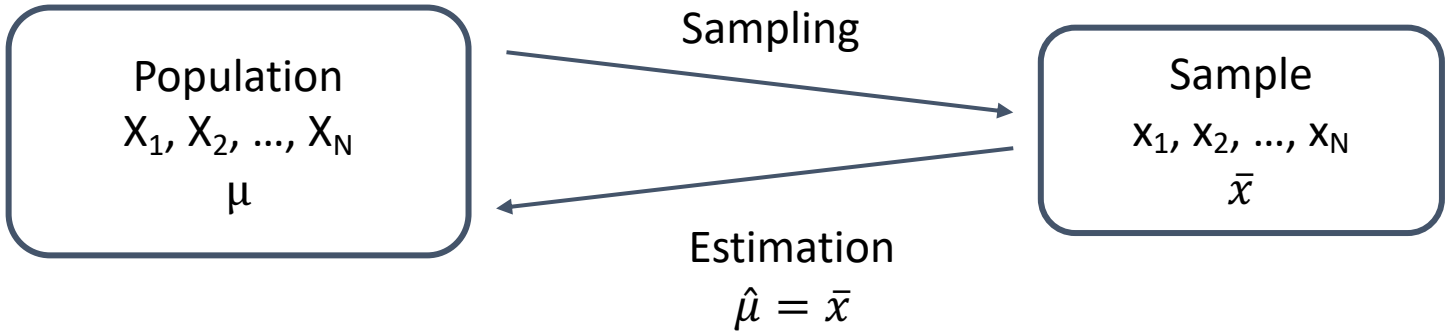
$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

where n in the sample size

Point estimates

Using sample data to calculate a single value which is “best guess” of an unknown population parameter.

Example: the population mean



Good characteristics of a point estimator

- Consistency

  The point estimator stays closer to the value of the parameter as we increase in size

- Unbiased

  The expected value of the point estimator is equal to the population parameter. $E(\bar{x}) = \mu$

- Sufficiency

  The sufficiency refers to how well an estimator utilized the information in the sample relative to eh postulated statistical model

Confidence interval

The confidence interval (CI) refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

Example: 95% confidence interval for population means

→ If we sampling for 100 times

→ Calculate 100 intervals

→ 95% of all intervals contain the real population mean

Hypothesis testing

Statistical hypothesis testing

A decision-making process with the help of statistical tools.

- Null hypothesis (H_0)
  Mostly what we are not interested in
- Alternative hypothesis (H_1 or H_a)
  Mostly the assumption which we wanted to prove

Proof by contradiction

Proportion:

Prove that H_1 is true

Method:

1. Assume that H_0 is true
2. Reject the H_0

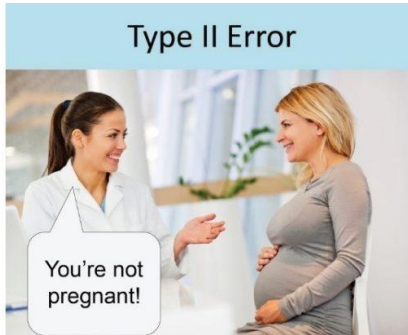
Possible conclusion:

- Reject H_0: H_1 is proven true
- Fail to reject H_0: We don't have enough evidence to reject H_0

Type I and Type II error

		Truth	
		H0 is true (non-carrier)	H1 is true (carrier)
Decision	Fail to reject H0 (negative)	Correct decision (true negative)	Type II error (false negative)
	Reject H0 (positive)	Type I error (false positive)	Correct decision (true positive)

If we wanted to test the newly developed rapid test kits.
→ Patient A tested positive while she is not a carrier.
→ Patient B tested negative while he is a carrier.



Type I and Type II error

		Truth	
		H0 is true (non-carrier)	H1 is true (carrier)
Decision	Fail to reject H0 (negative)	Correct decision (true negative)	Type II error (false negative)
	Reject H0 (positive)	Type I error (false positive)	Correct decision (true positive)

- $\alpha = \mathrm{P}(\text{Type I error})$: Significant level
- $\beta = P(\text{Type II error})$
  - Statistical power (H_1 is true and H_0 us rejected)

    ➔ Power = $1 - \beta$

Steps of hypothesis testing

[Example]

A teacher claims that students in her school are smarter than others.

- Sample size = 30
- Average IQ score = 112.5
- Population SD = 15
- Population average = 100

[Step 1]

State the hypothesis

(alternative → null hypothesis)

H_1: $\mu > 100$

H_0: $\mu \leq 100$

** NOTE: the H_0 must have "="

[Step 2]

Set significant level. $\alpha = 0.05$

Steps of hypothesis testing

[Example]

A teacher claims that students in her school are smarter than others.

- Sample size = 30
- Average IQ score = 112.5
- Population SD = 15
- Population average = 100

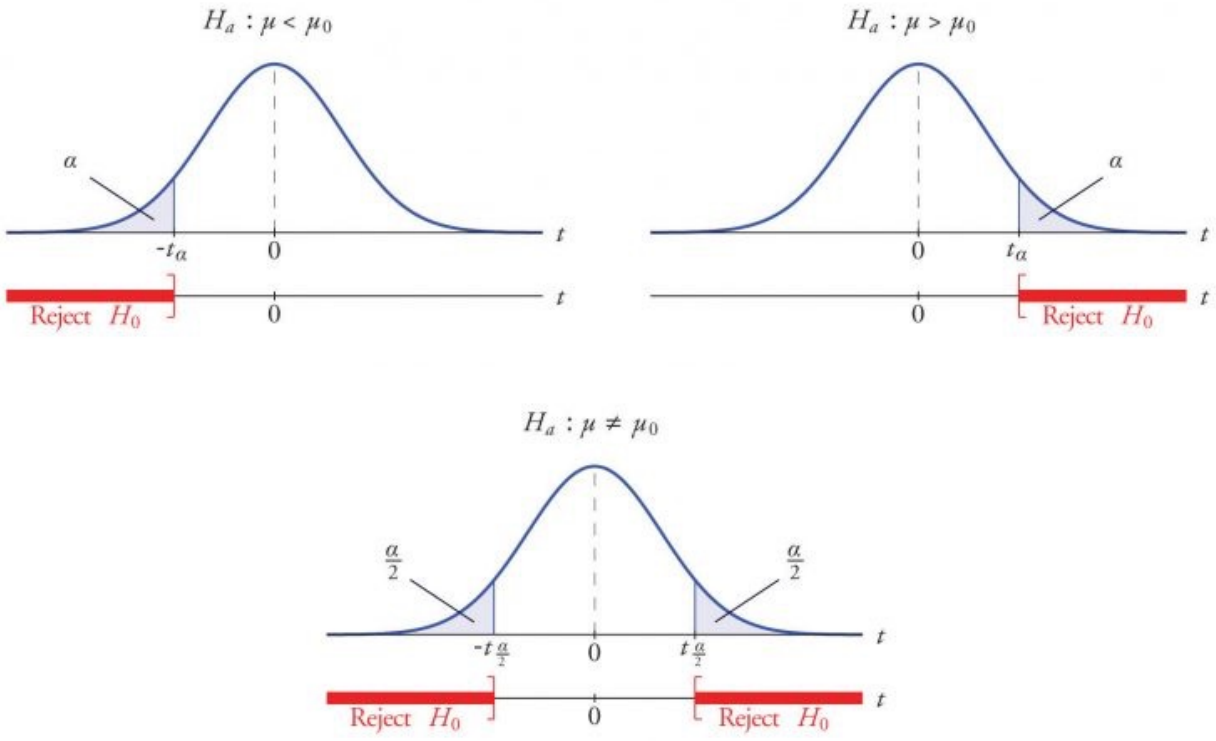
[Step 3]

Calculate the test statistics and/or p-value

[Step 4]

Make decision

Decision making

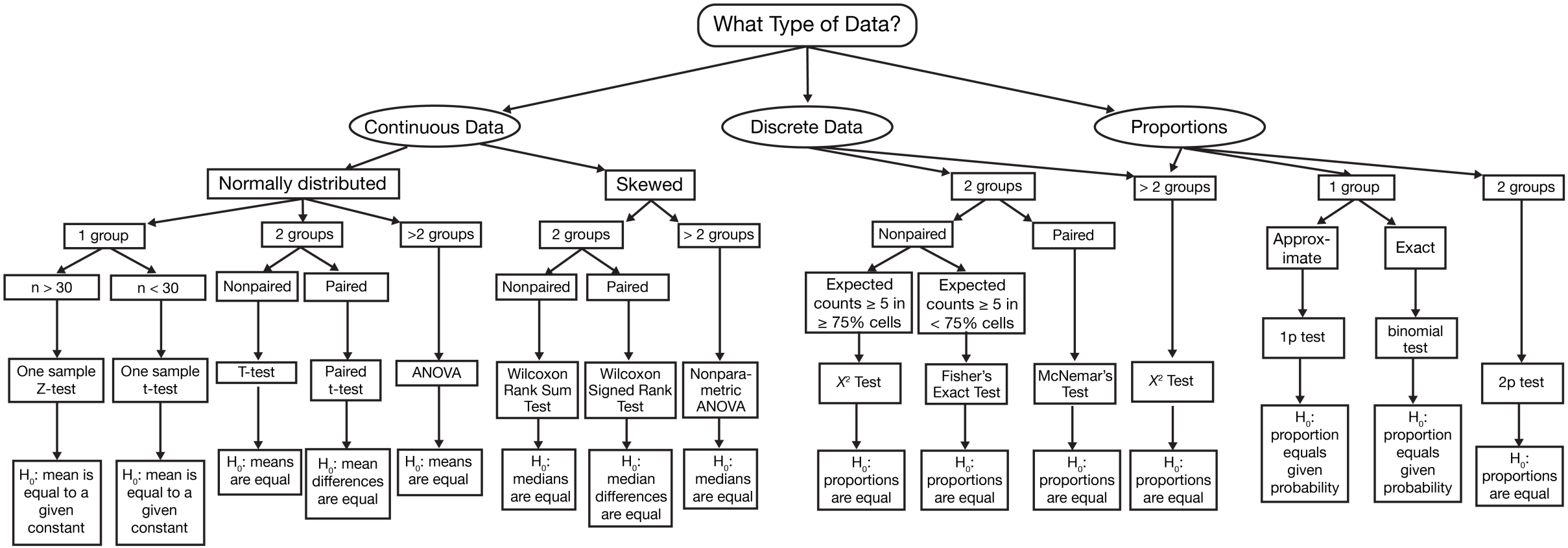


The plot shows reject regions of three different hypotheses. The idea is that the farther the observation and the mean are, the less likely they are from the same distribution.

Three ways to make the decision

- Convert the observed value to Z
- Convert the border Z to realistic distribution
- Calculate p-value

Flow chart: which test statistic should you use?



Z test

Aim: Test if the weights of milk powder are less than 500g.

- Random sampled (n = 36)
- Sample mean $\bar{x} = 485\text{g}$
- Population SD = 30g

[STEP 1] H_0: $\mu \geq 500$ vs. H_1: $\mu < 500$

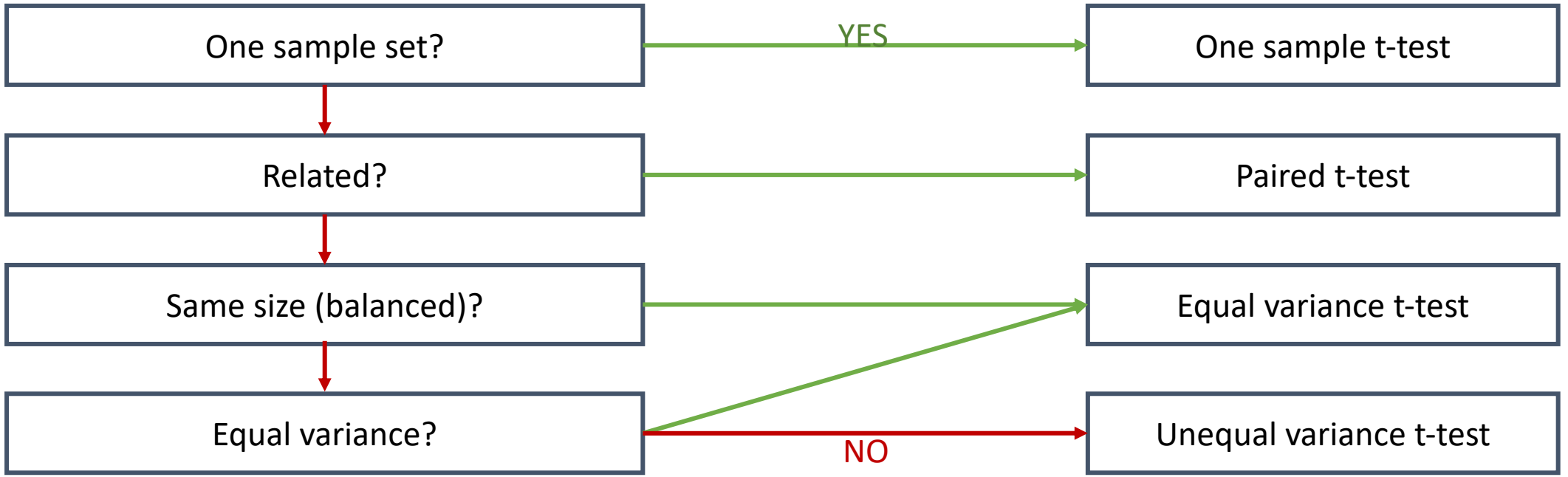
[STEP 2] $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{485-500}{30/\sqrt{36}} = -3$

[STEP 3] Z = -3 < -1.645 = $Z_{0.05}$

[STEP 4] Reject H_0.

Student t-test

- Compare means
- Four different types of t-test:



One sample t-test and paired t-test

One sample t-test

Test statistics

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Paired t-test

$$D = x_1 - x_2$$

Test statistics

$$t = \frac{\overline{D}}{s_D/\sqrt{n}}$$

Two sample t-test

Equal variance

Test statistics

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{S_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

where

$$S_p = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

Degrees of freedom: $n_A + n_B - 1$

Unequal variance

Test statistics

$$t' = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Degrees of freedom

$$df' = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{s_A^2/n_A}{n_A - 1} + \frac{s_B^2/n_B}{n_B - 1}}$$

t-test function in R

```
t.test(x, y = NULL,
     alternative = c(“two.sided”, “less”, “greater”),
     mu = 0, paired = FALSE, var.equal = FALSE,
     conf.level = 0.95)
```

Result from `t.test()` function

```
Two Sample t-test
data: x and y
t = -3.2744, df = 98, p-value = 0.001464
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.201314 -1.275845
sample estimates:
mean of x mean of y
30.94980 34.18838
```

Fisher's exact test

Fisher's exact test is used in the analysis of the contingency table. We want to know whether any difference in proportions that we observed is significant. (Especially used when the sample size is small)

H_0: The proportion of catching different rabbit variety on both traps are the same

	Trap A	Trap B	Total
Rabbit A	4 (a)	1 (b)	5 (a+b)
Rabbit B	3 (c)	4 (d)	7 (c+d)
Total	7 (a+c)	5 (b+d)	12 (a+b+c+d)

Fisher’s exact test

	Trap A	Trap B	Total
Rabbit A	4 (a)	1 (b)	5 (a+b)
Rabbit B	3 (c)	4 (d)	7 (c+d)
Total	7 (a+c)	5 (b+d)	12 (n = a+b+c+d)

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

```
fisher.test(x = TABLE, alternative, conf.level = 0.95)
```

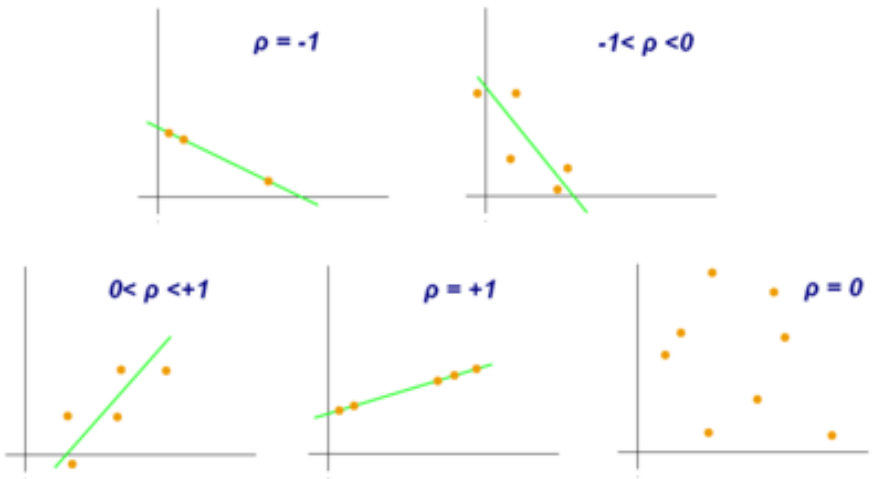
Result of Fisher’s exact test

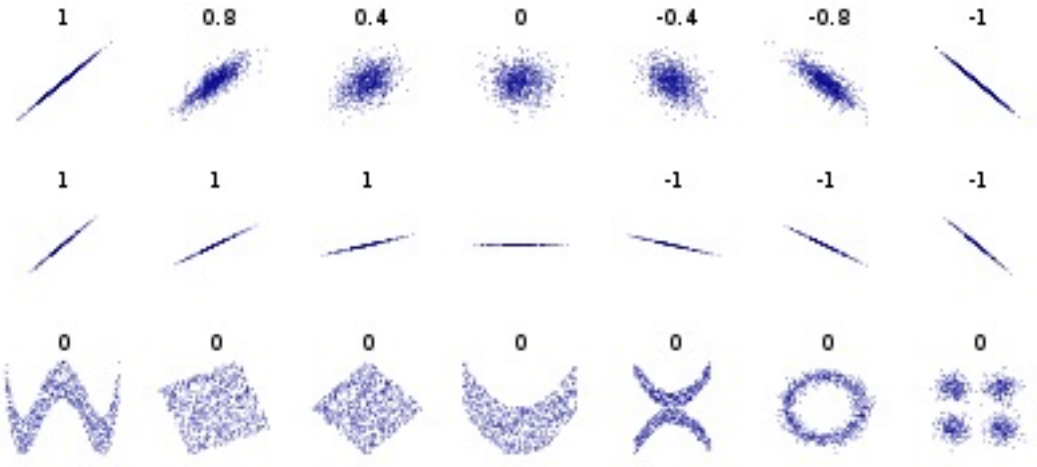
```
        Fisher's Exact Test for Count Data
data: rabbit.trap
p-value = 0.2929
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
     0.2536 326.1073
sample estimates:
odds ratio
     4.609061
```

Correlation coefficient

- Relationship between two random variables
- The most familiar measure is the Pearson's correlation coefficient
- For two random variables X and Y

$$\rho_{XY} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$





Correlation coefficient

cor(x, y)

cor.test(x, y)

```
Pearson's product-moment correlation
data: x and y t = 0.041508, df = 48, p-value = 0.9671
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
      -0.2728116 0.2838654
sample estimates:
      cor
      0.005991052
```

Simple linear regression

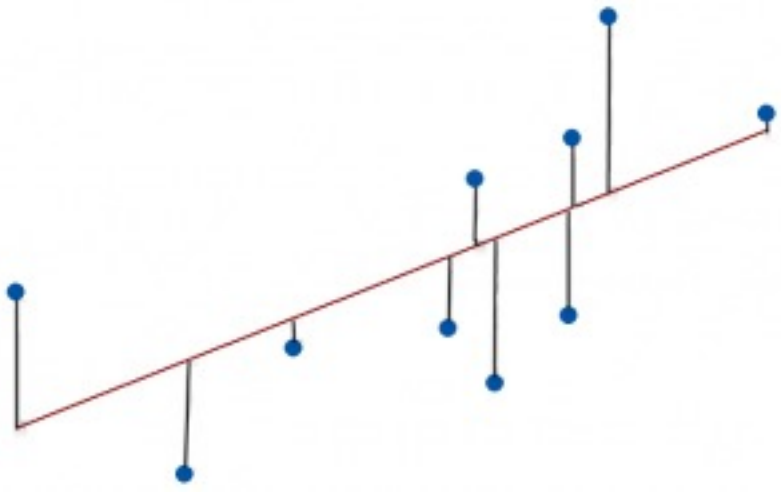
Linear regression is a linear approach to modeling the relationship between dependent and independent variables.

$$y = X\beta + \epsilon$$

There are a few ways to determine the best-fitted model, one of them is the least-squares method.

➔ Error term $\epsilon = y - X\beta$

➔ The best fit is the line that has the smallest $\epsilon^T \epsilon$



Fitting linear model in R

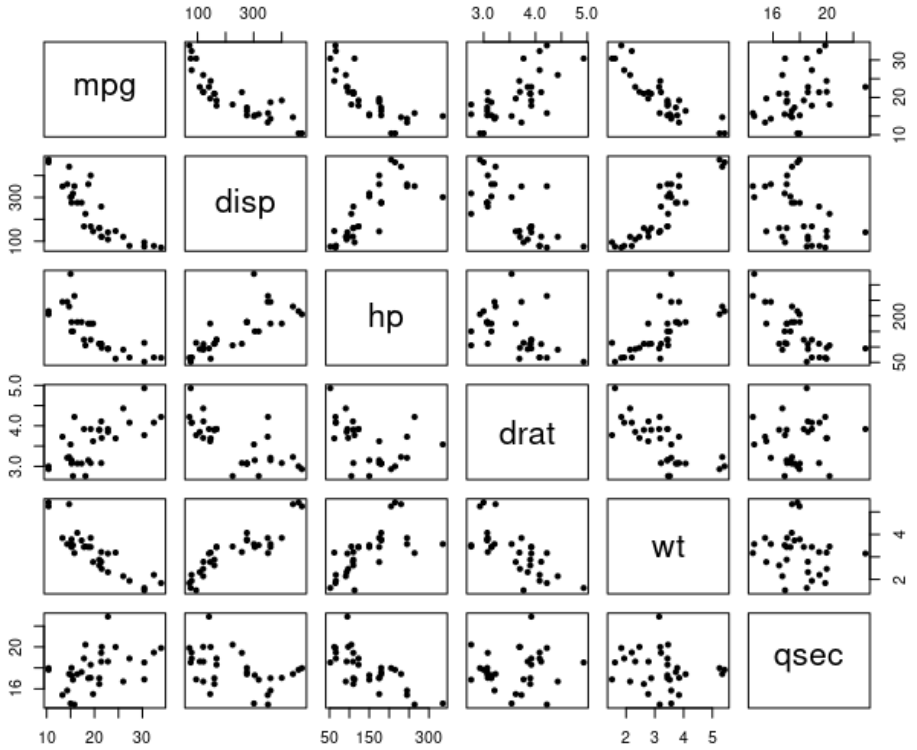
```
fit = lm(formula, data)
```

- “=” should be replaced by “~” when writing formula in R since “=” has been used as assigned
- Example: $y = b_0 + x_1 b_1 + x_2 b_2$
  Could be written as formula = $y \sim x_1 + x_2$
- Use `summary()` function to return the result of `lm()`
  `summary(fit)`

Try on an example data: data("mtcars")

- 1974 from the Motor Trend US magazine
- Extract columns 1, 3, 4, 5, 6, 7 for this practice

	mpg	disp	hp	drat	wt	qsec
Mazda RX4	21.0	160.0	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	160.0	110	3.90	2.875	17.02
Datsun 710	22.8	108.0	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	258.0	110	3.08	3.215	19.44
Hornet Sportabout	18.7	360.0	175	3.15	3.440	17.02
Valiant	18.1	225.0	105	2.76	3.460	20.22
Duster 360	14.3	360.0	245	3.21	3.570	15.84
Merc 240D	24.4	146.7	62	3.69	3.190	20.00
Merc 230	22.8	140.8	95	3.92	3.150	22.90



Why should we check correlations? Collinearity!

The result from the `lm()` function

```
fit = lm(mpg ~ wt, data = mtcars)
summary(fit)
```

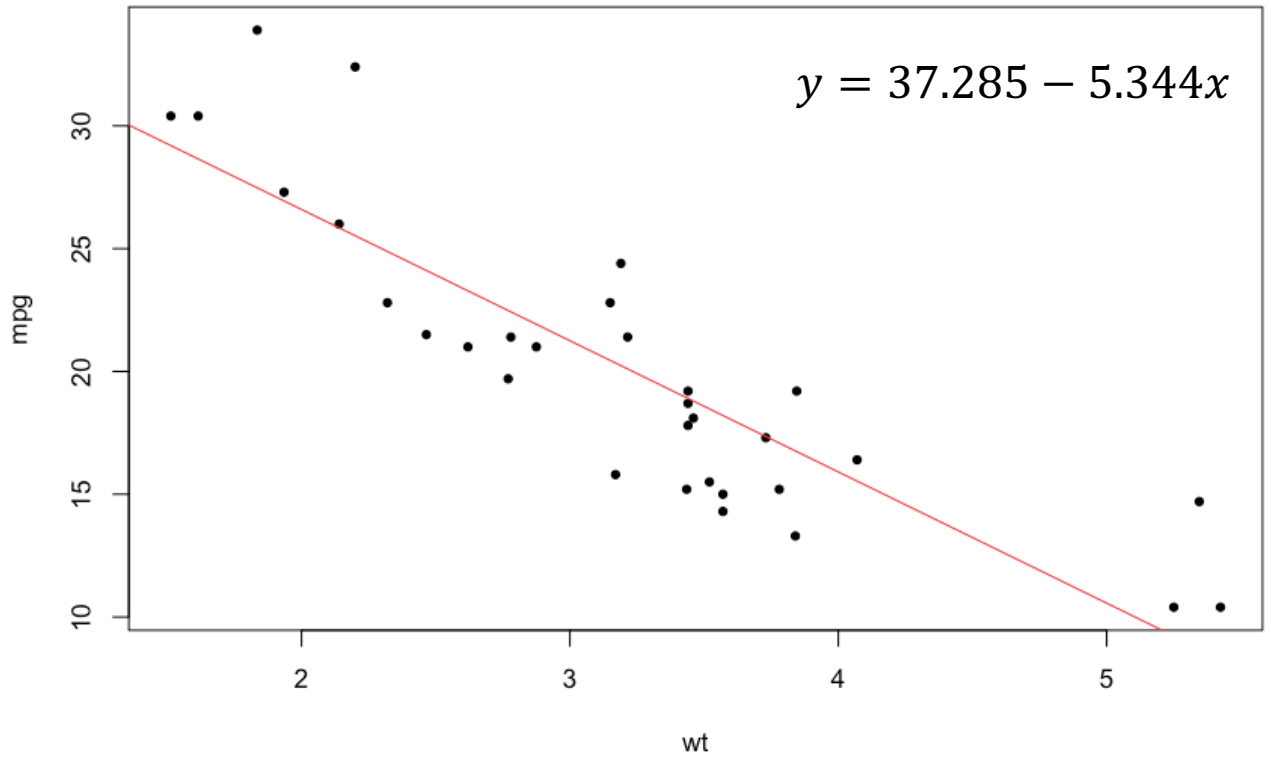
```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Scatter plot and the fitted model

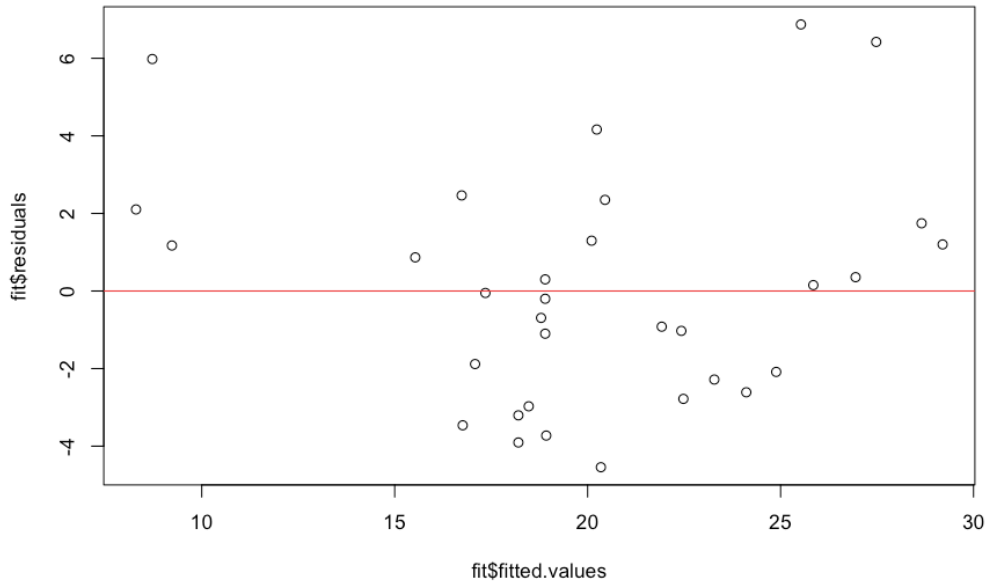


The plot of residuals vs. fitted value

As we'd learned, the relationship between x and y can be described by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

If y have perfectly described by x, the error e should be evenly distributed around 0. Thus, a residual plot vs. fitted value is a good tool for checking this.



Analysis of variance (ANOVA)

- Based on the law of total variance
- Determine whether there are any statistically significant differences between the means of 3 or more independent groups.

```
fit = aov(formula, data)
summary(fit)
```

Practice