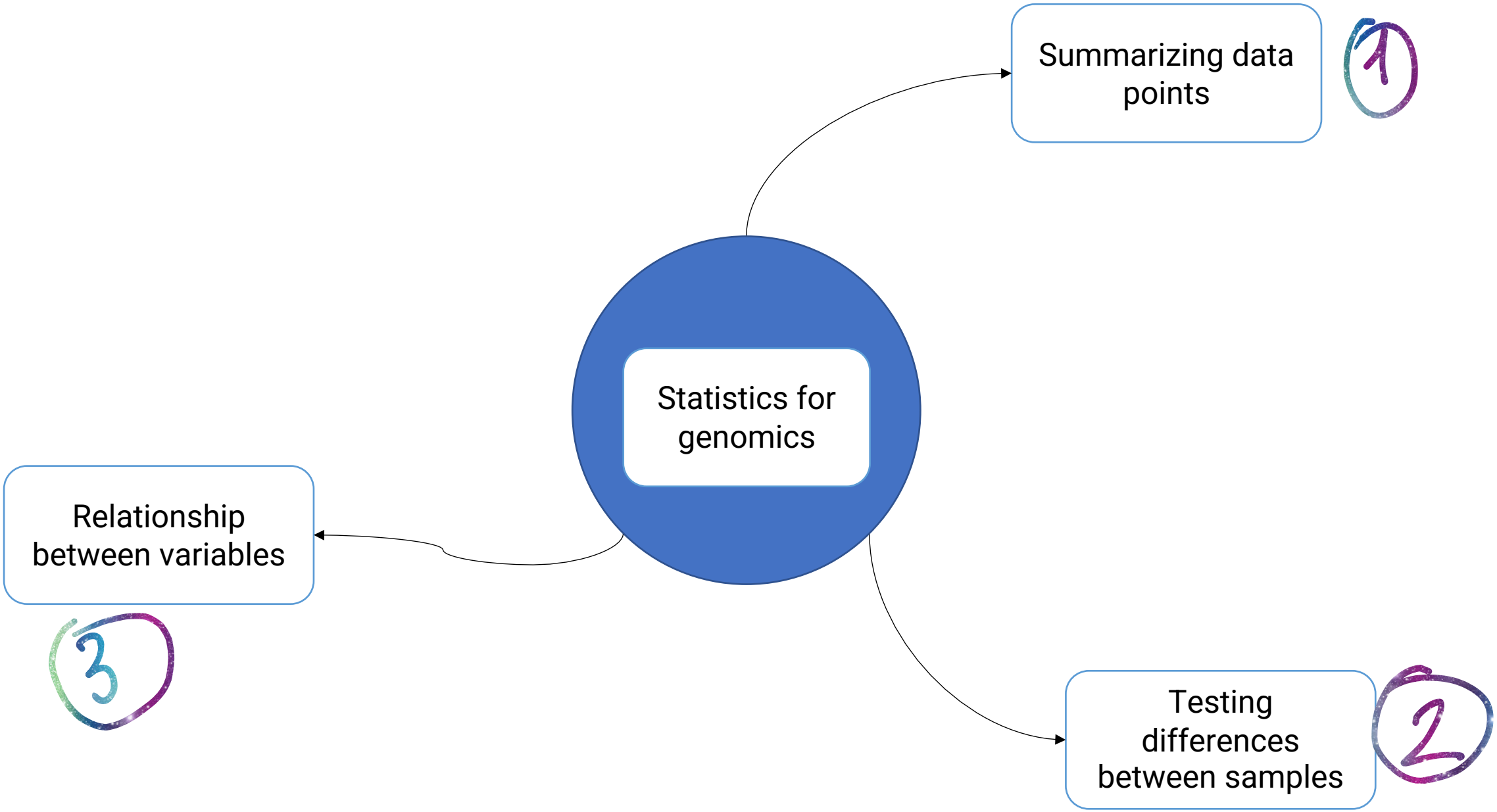
Computational genomics: hands on course

Statistics for genomics


Summarizing data points
1
Statistics for genomics
Relationship between variables
3
Testing differences between samples
2

How to summarize collection of data points:

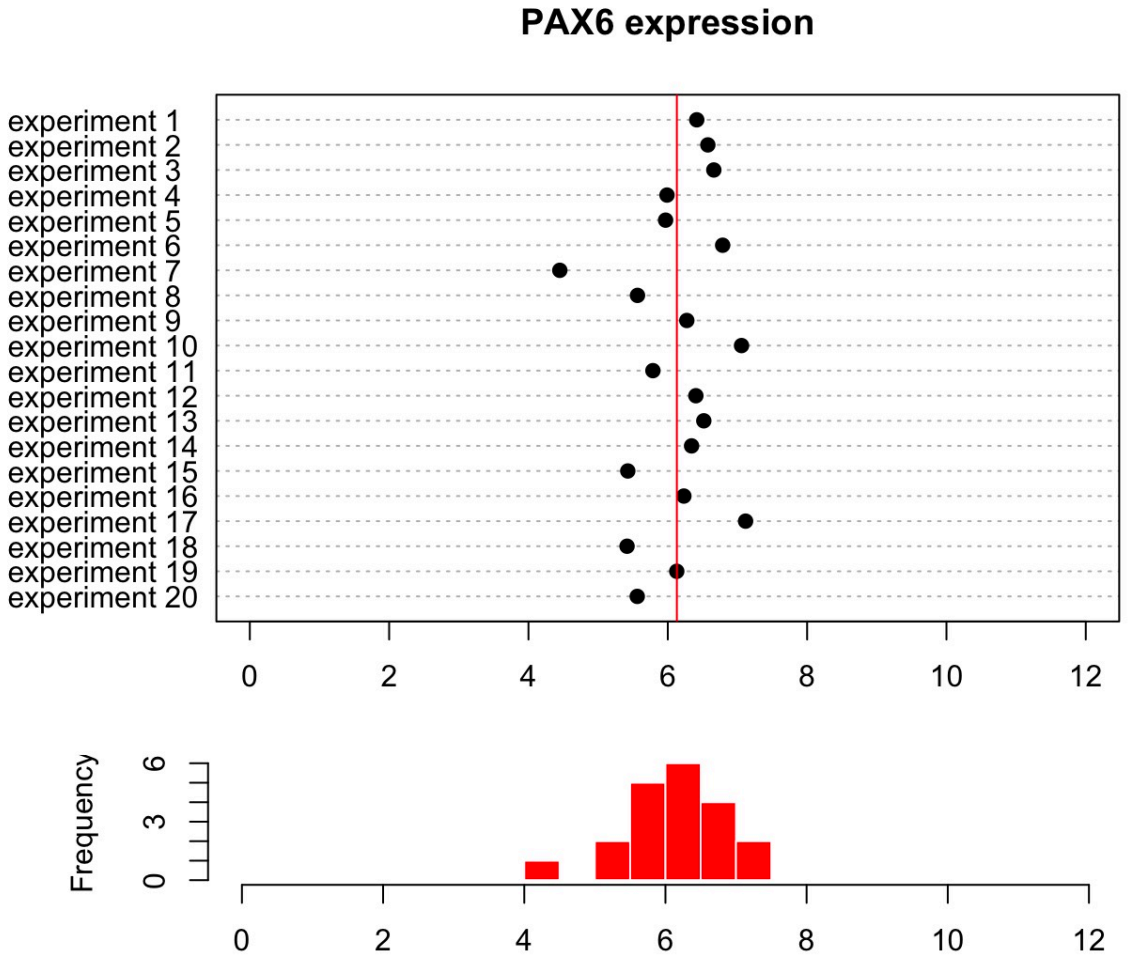
The idea behind statistical distributions

37.4

37.6
37.7
37.5
37.8

How to summarize collection of data points:

The idea behind statistical distributions



general points

How to summarize collection of data points:

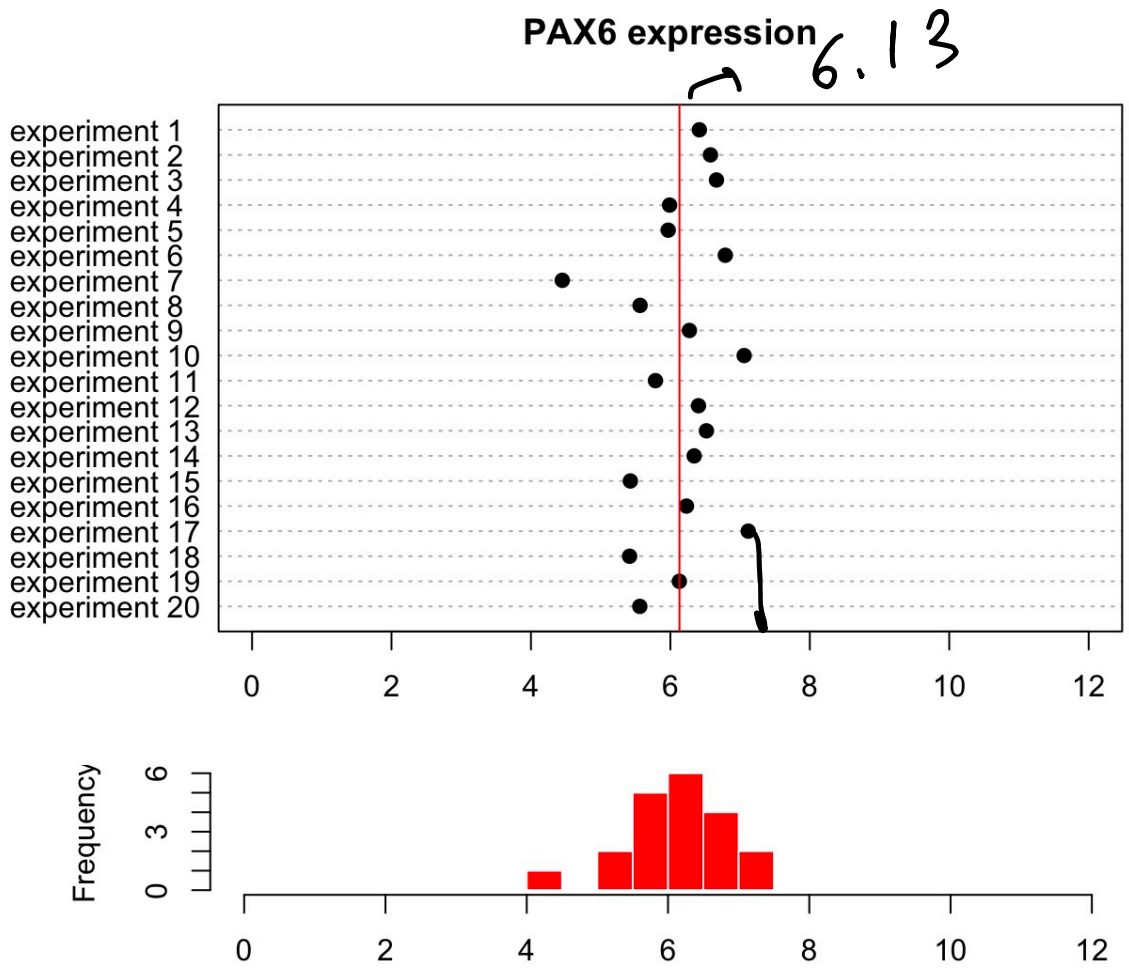
The idea behind statistical distributions

- Experimentation -> data
- Each data point is different
  - Biological variation or measurement error is to blame

How to describe a collection of data points

- What is the most frequent value ?
- How variable are your measurements ?

Describing the central tendency:
mean and median

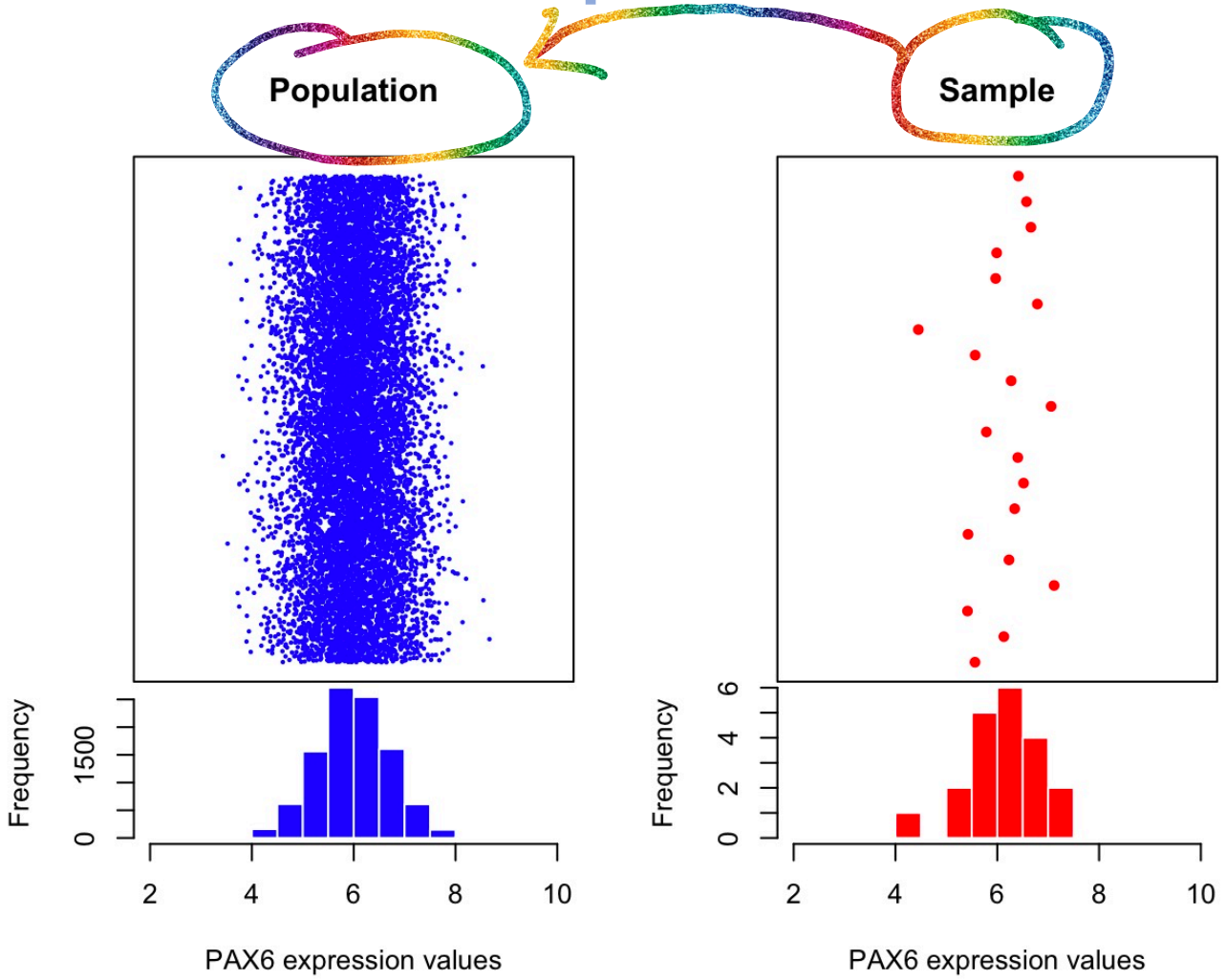


$$\overline{X} = \sum_{i=1}^{n} x_i / n$$

$$\frac{4 + 4.5 - - + 7.2}{20}$$

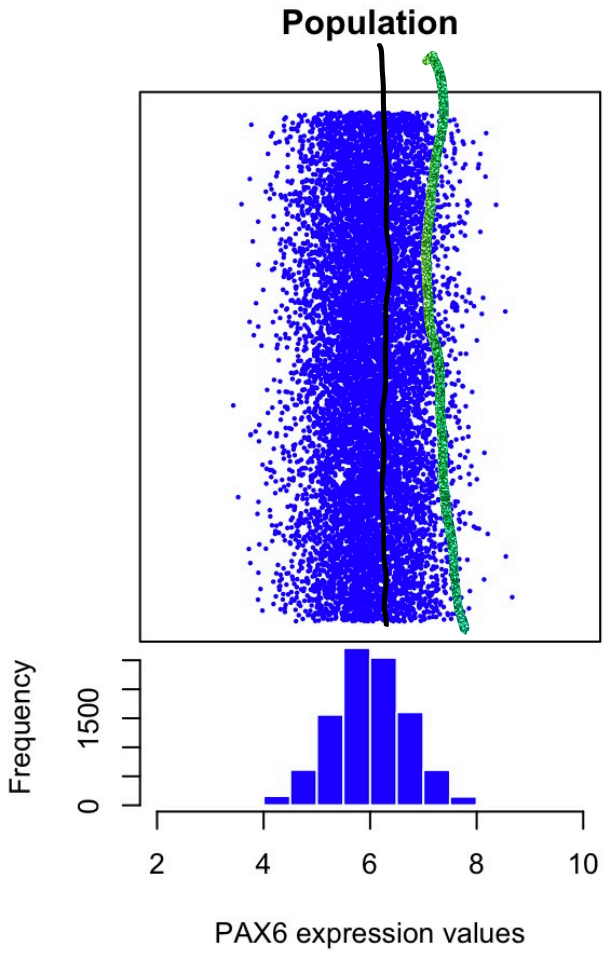
$$= 6.13$$

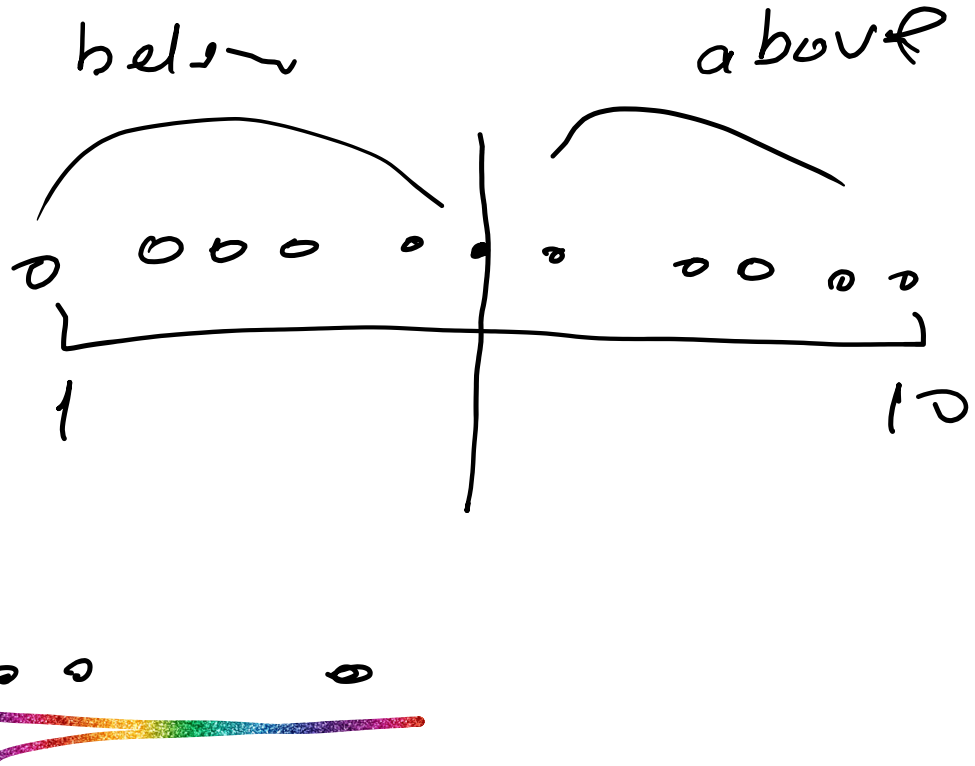
Describing the central tendency:
population -> sample



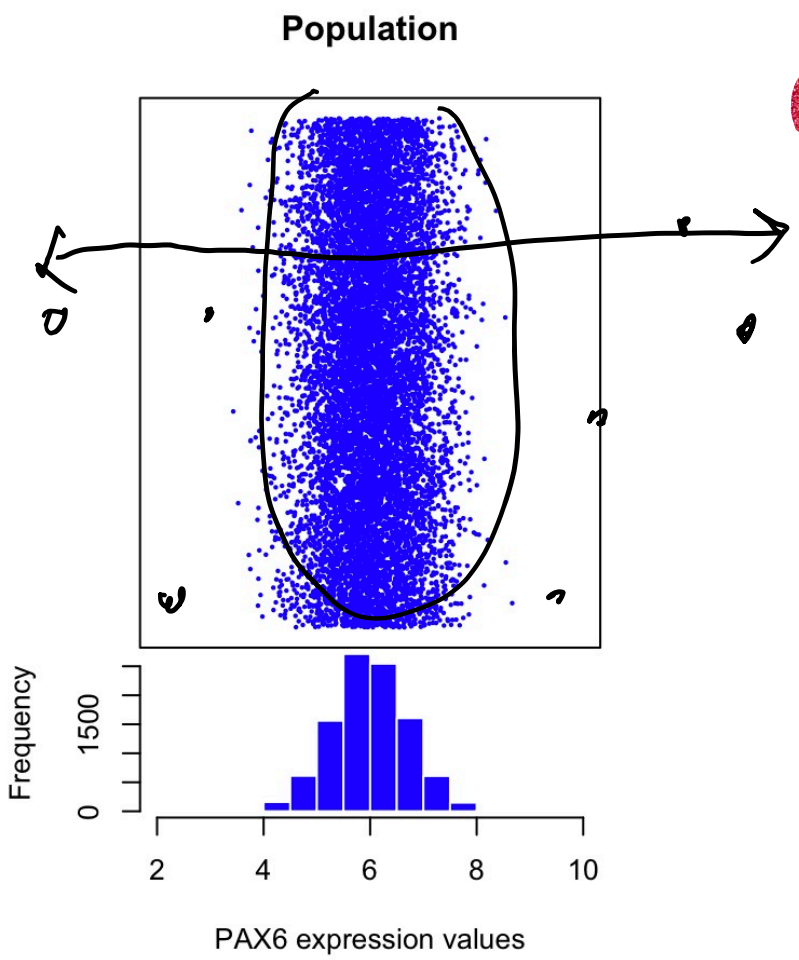
Describing the central tendency:

median



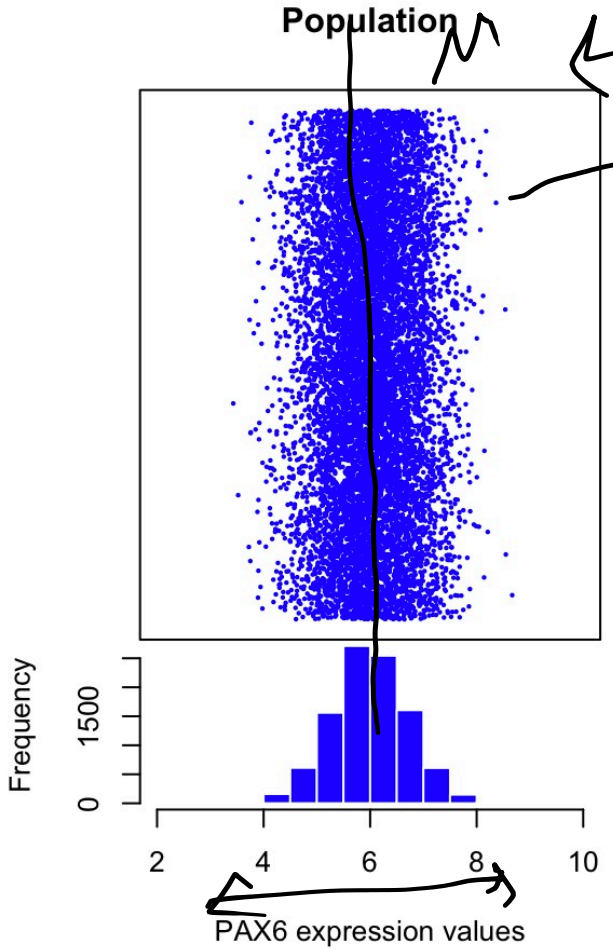


Describing the spread:
measurements of variation



range min — max

Describing the spread:
measurements of variation



$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

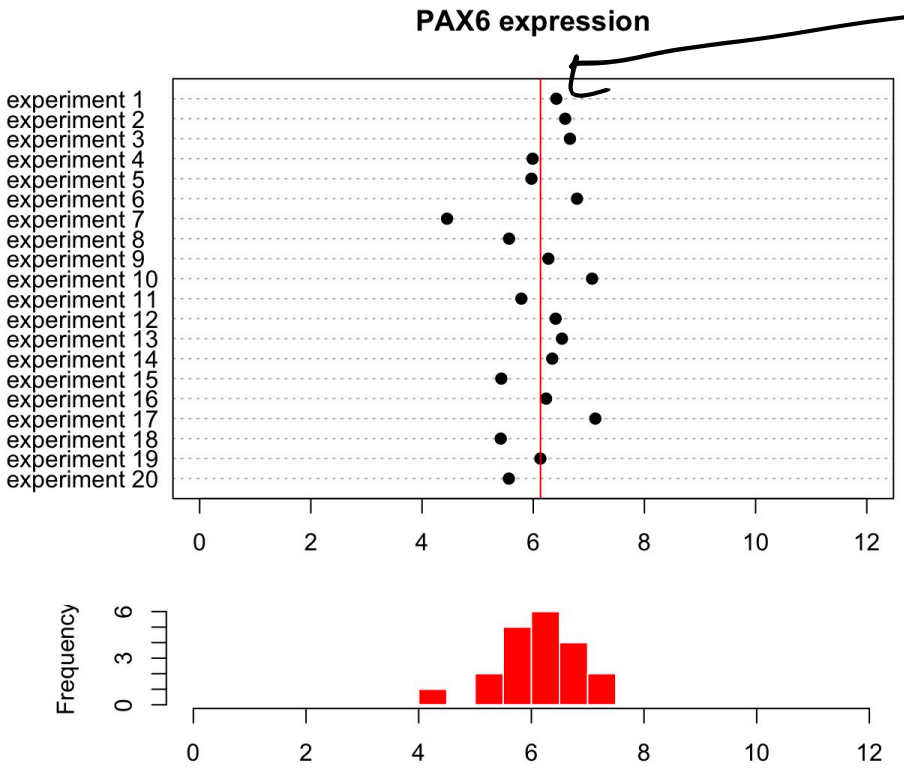
variance

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

std. dev.

points

Describing the spread:
measurements of variation

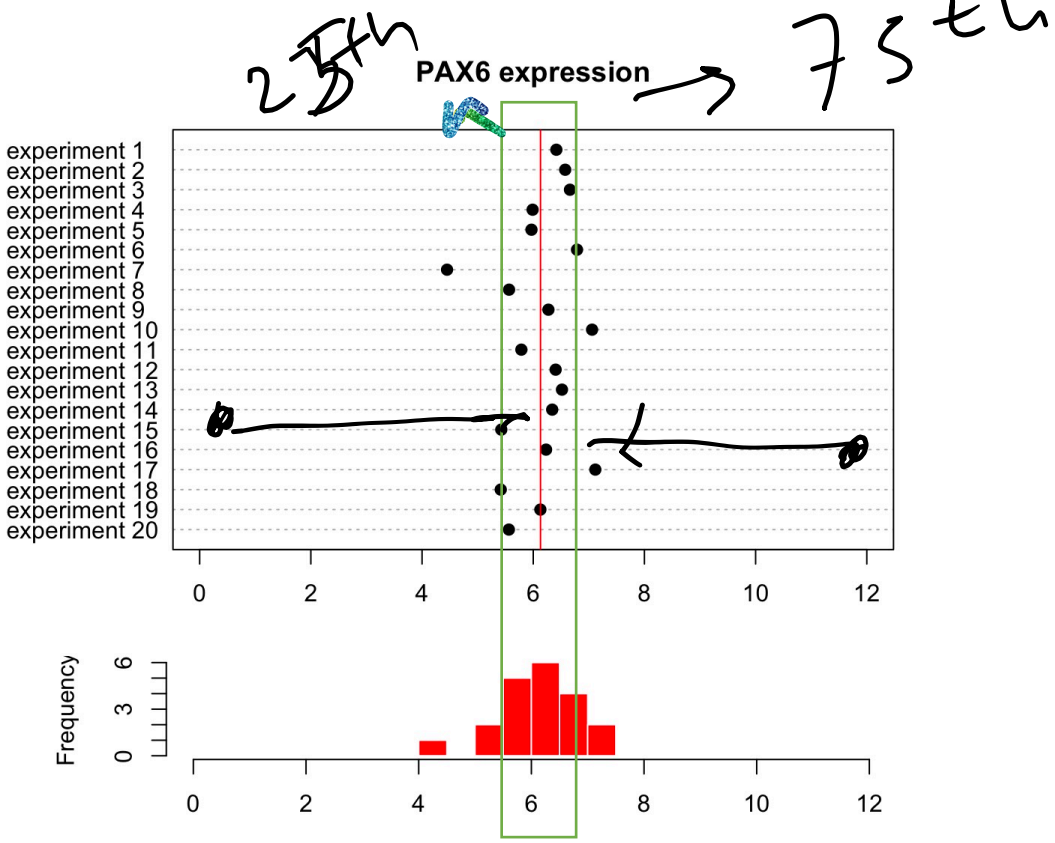


sample mean

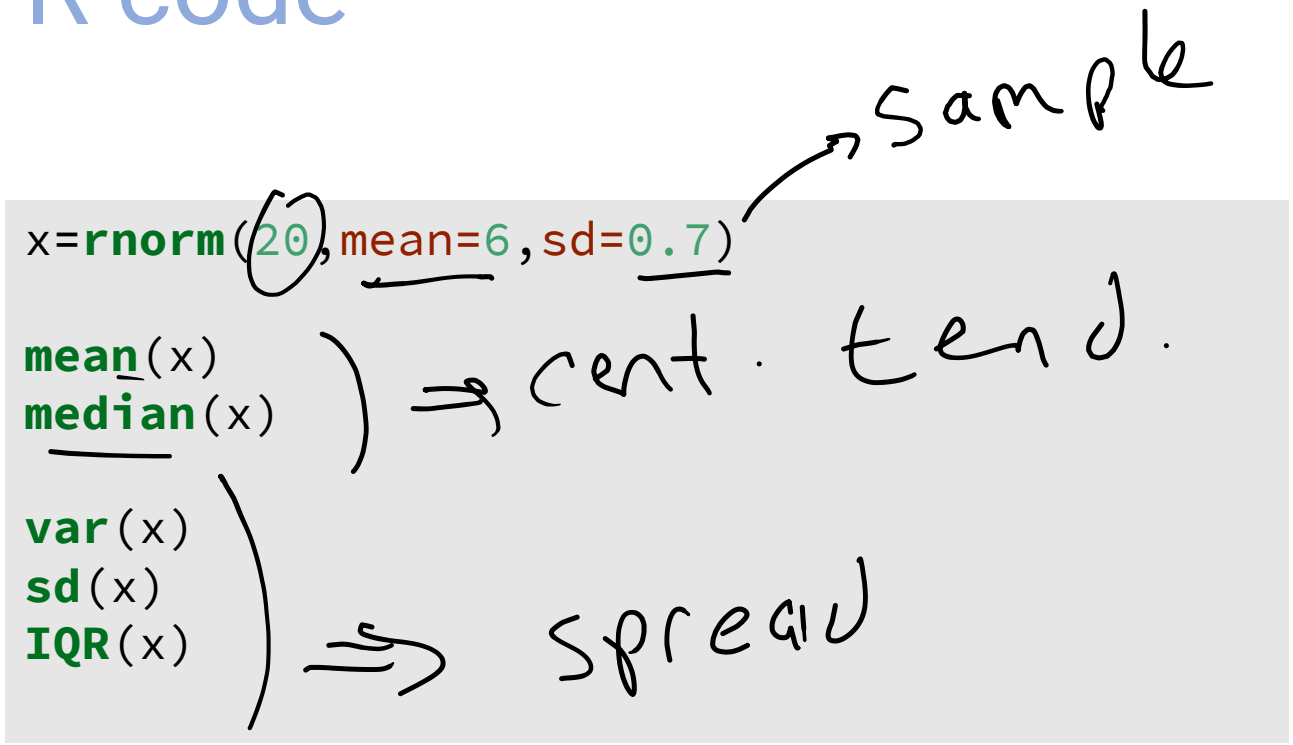
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}$$

points - 1

Describing the spread:
interquartile range

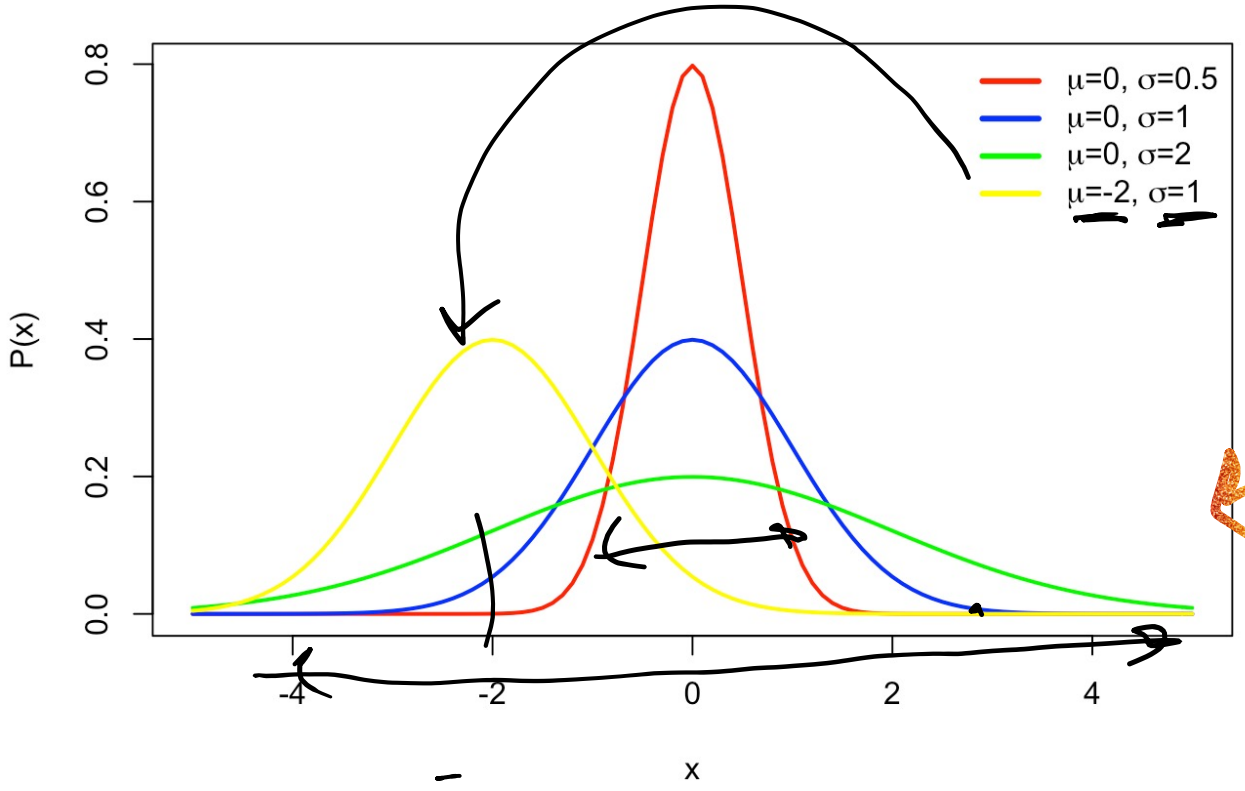


Describing the spread and central tendency:
R code



Frequently used statistical distributions

Normal distribution



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

mean

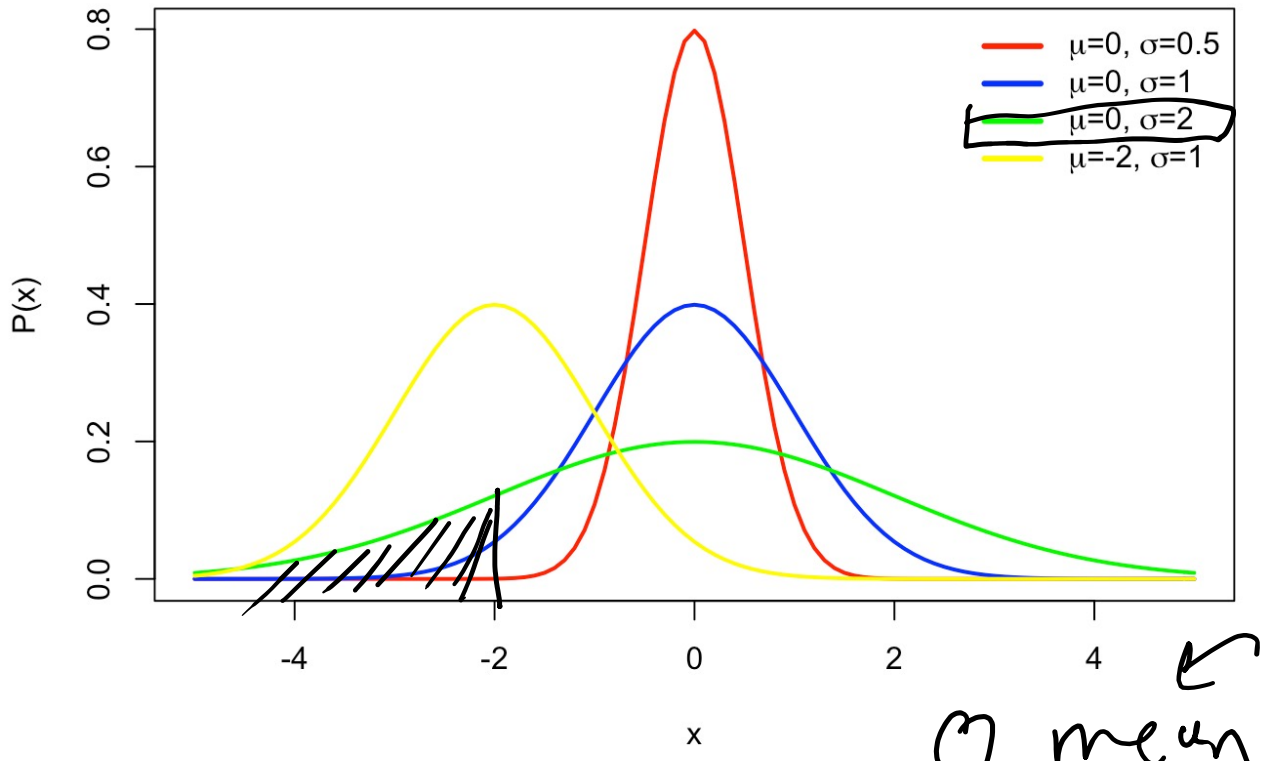
var.

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

density

Frequently used statistical distributions

Normal distribution



$$P(X <= -2 \mid \mu = 0, \sigma = 2)$$

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

Std.

0 mean

1 std. dv

table

$Z \Rightarrow$ prob.

Frequently used statistical distributions

Normal distribution

$$P(X <= -2 \mid \mu = 0, \sigma = 2)$$

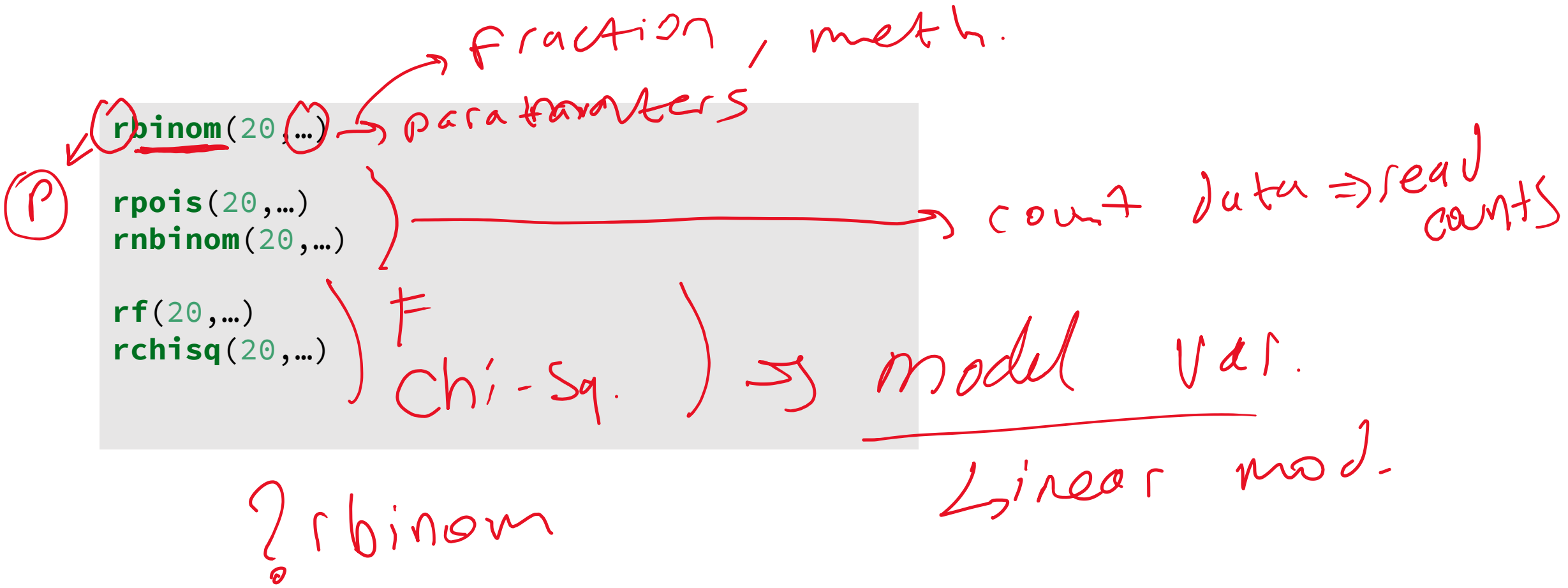
```r
#get the probability of P(X =< -2) | mean=0 and sd=2
pnorm(-2, mean=0, sd=2)

#get the probability of P(X > -2) | mean=0 and sd=2
pnorm(-2, mean=0, sd=2,lower.tail = FALSE)

#get 5 random numbers from norm. dist. | mean=0 and sd=2
rnorm(5, mean=0 , sd=2)
```

Frequently used statistical distributions

Other distributions: R code

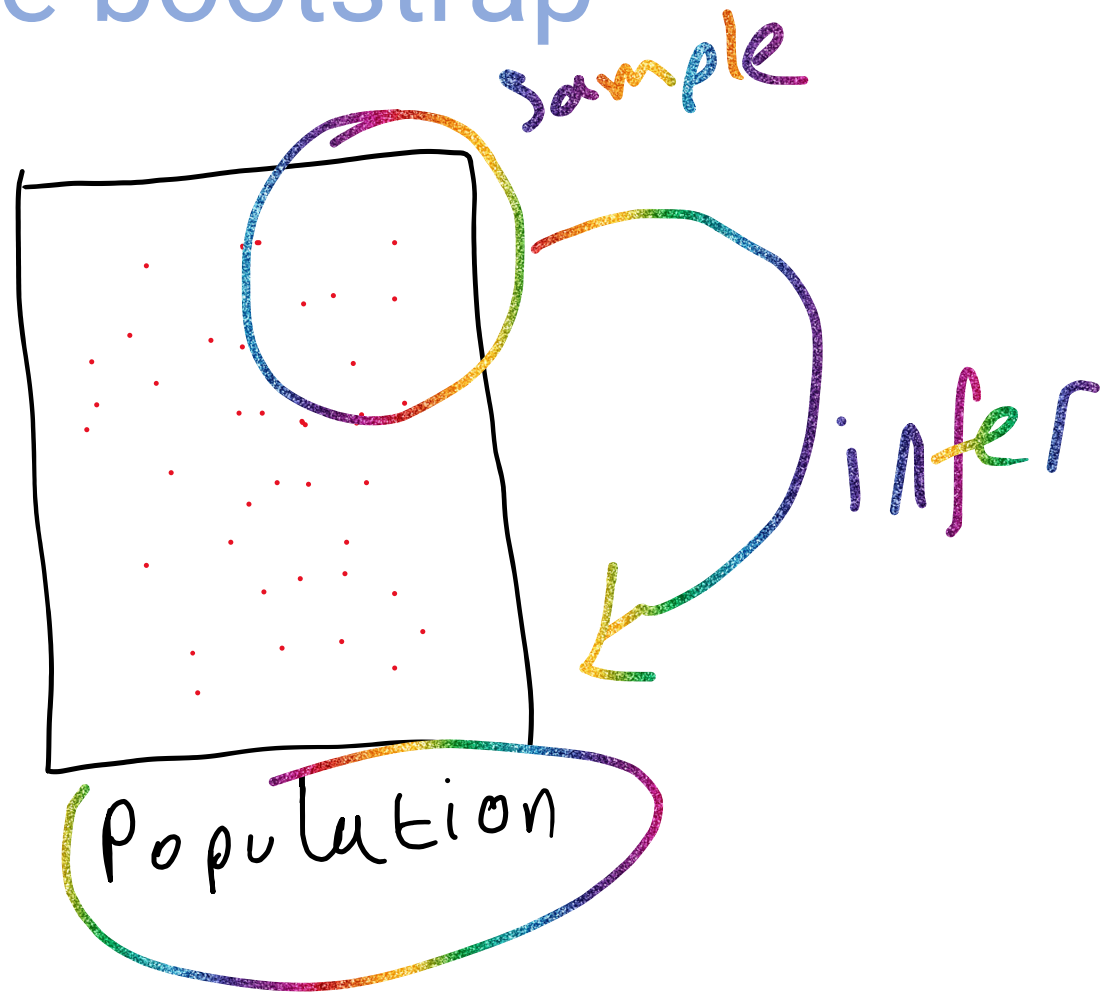


Precision of estimates:
confidence intervals

- How well a statistic estimated from a sample represents the population statistic ?

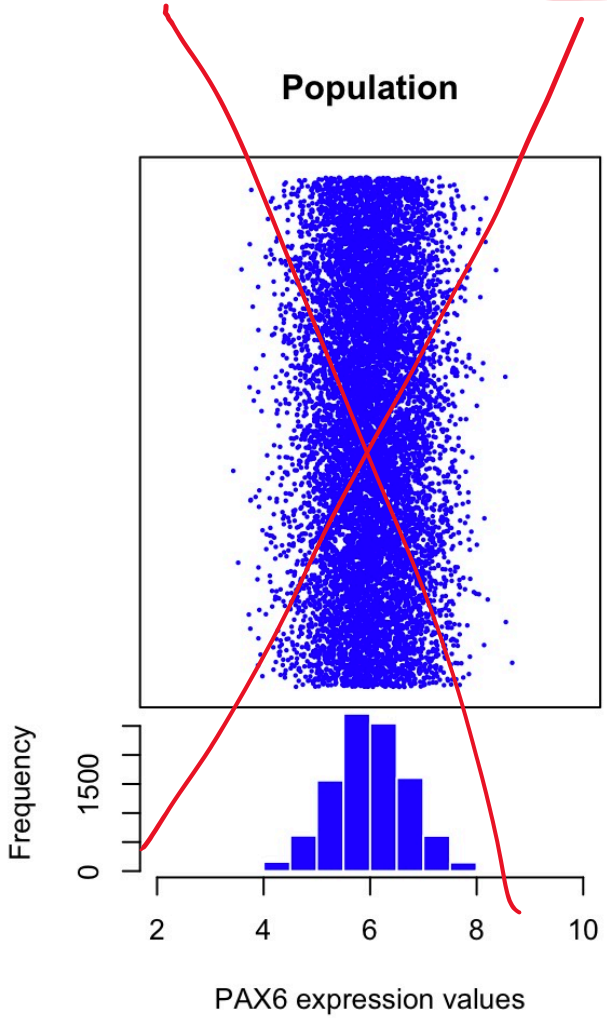
→ range for stat.
mean
std. dev.

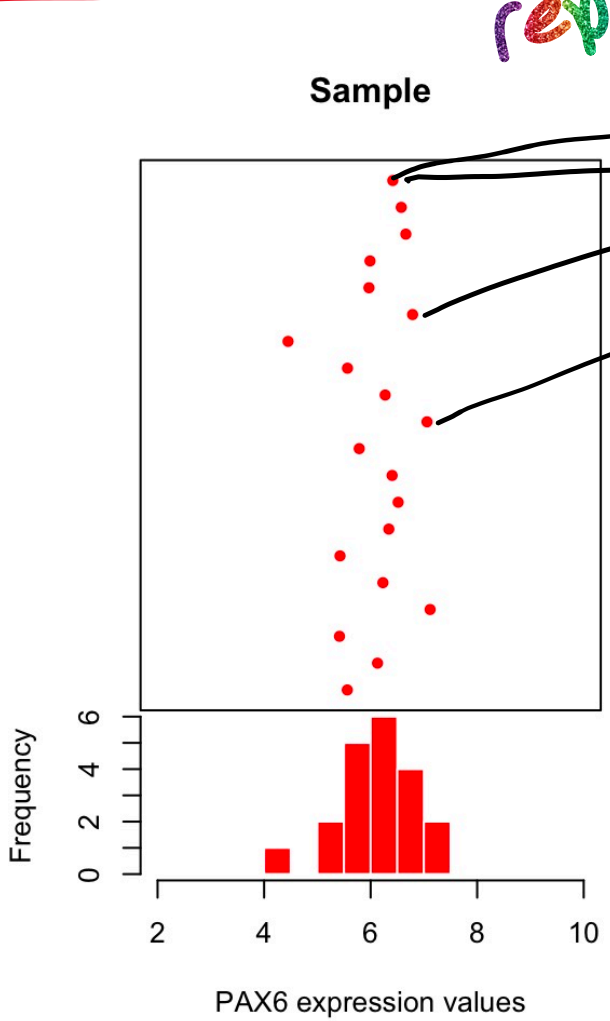
Precision of estimates:
The bootstrap



Precision of estimates:

The bootstrap





with replacement

n=20

1 → $\bar{X}_1$

2 → $\bar{X}_2$

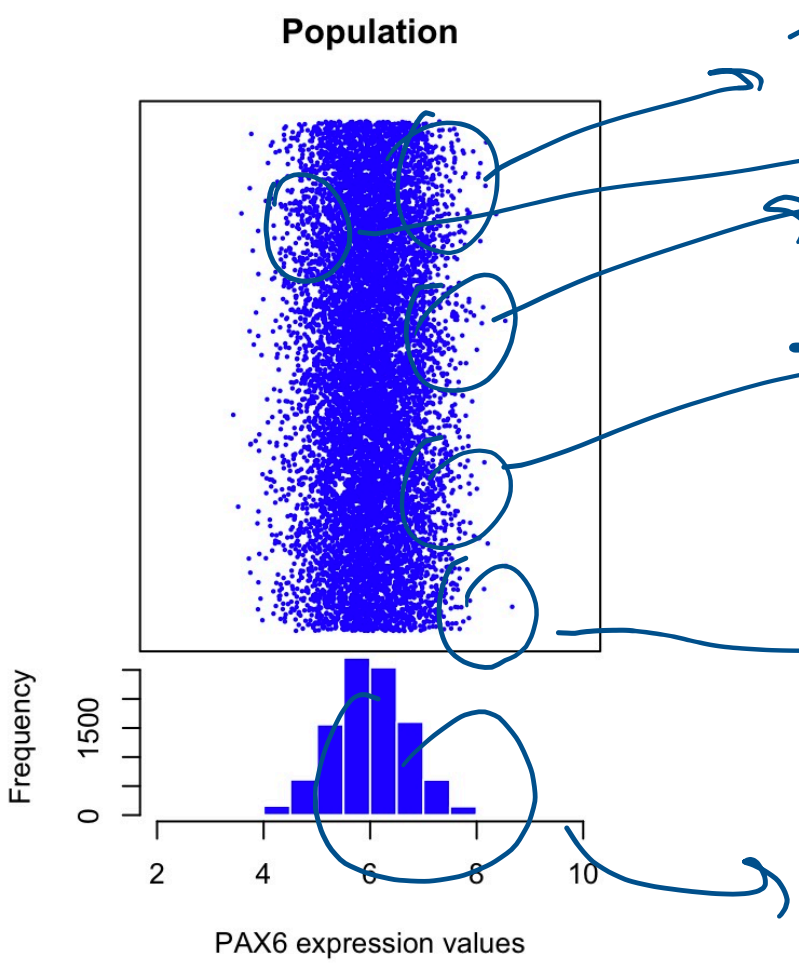
1000 → $\bar{X}_{1000}$

2.5th

97.5th

95% Confo Int.

Precision of estimates:
Central Limit Theorem



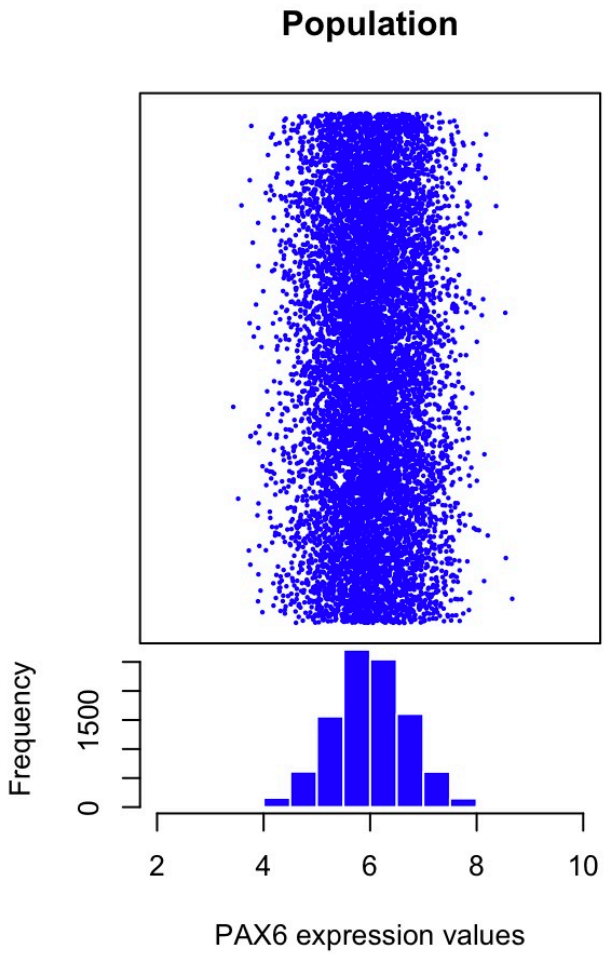
$\bar{X}_1$
$\bar{X}_2$
$\bar{X}_3$
$\bar{X}_4$
$\bar{X}_n$

normal
binomial
expo.
any dist.

95% Conf. int.
pop mean μ
$\sigma/\sqrt{n}$
pop std. dev. $\sqrt{n}$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Precision of estimates:
Central Limit Theorem



$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

$$P(-1.96 < Z < 1.96) = 0.95$$

$$P\left(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$P(\mu - 1.96\sigma/\sqrt{n} < \overline{X} < \mu + 1.96\sigma/\sqrt{n}) = 0.95$$

$$P(\overline{X} - 1.96\sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

$$confint = [\overline{X} - 1.96\sigma/\sqrt{n}, \overline{X} + 1.96\sigma/\sqrt{n}]$$

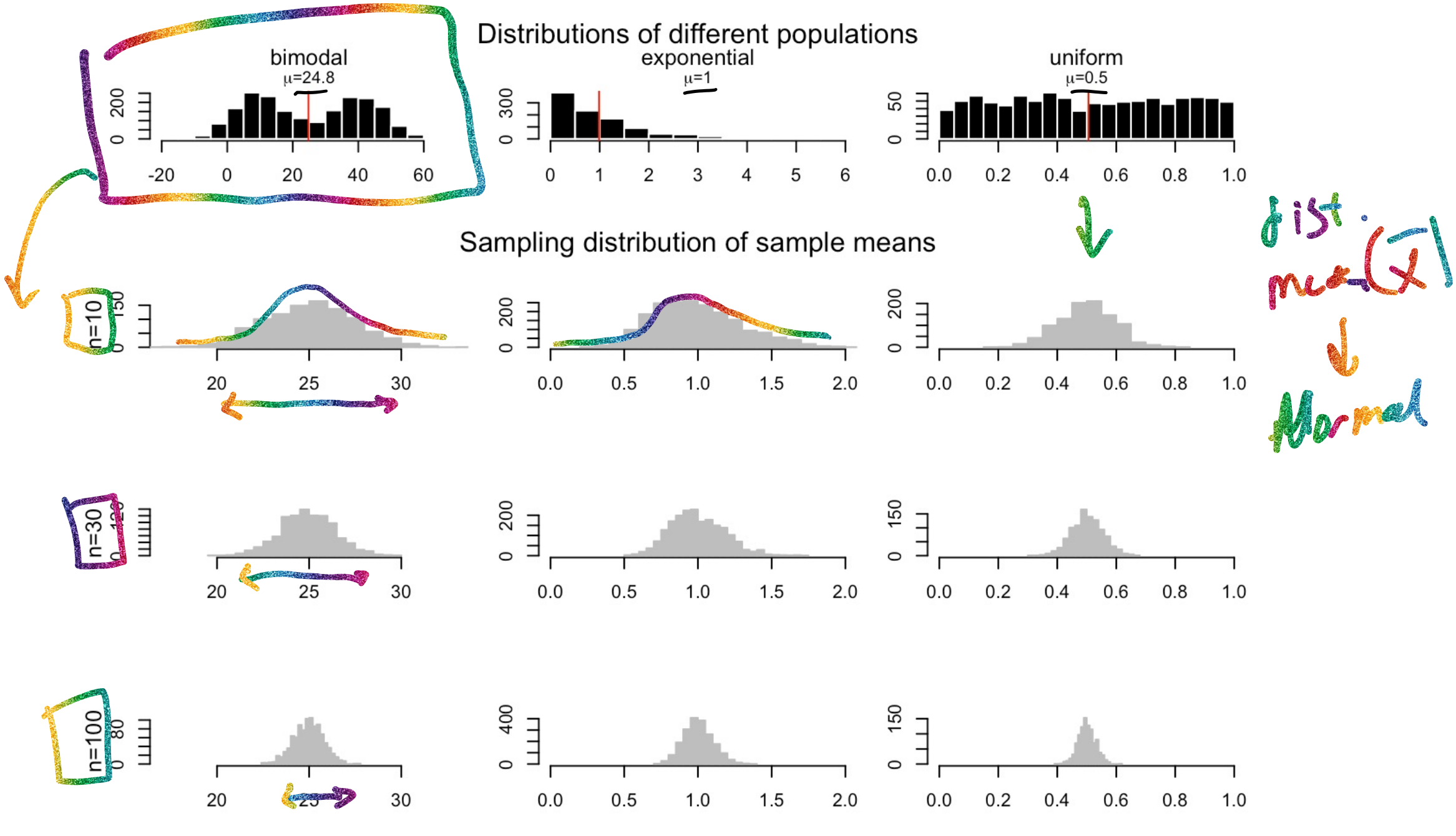
90%

$$\overline{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$$

$N(0,1)$

-1,96 +1,96

5^{th} , $95^{th} \Rightarrow N(0,1)$


Distributions of different populations
bimodal
μ=24.8
exponential
μ=1
uniform
μ=0.5
Sampling distribution of sample means
n=10
n=30
n=100
dist
mean(x1)
Normal

CLT problem: we need population std. dev.

$$\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

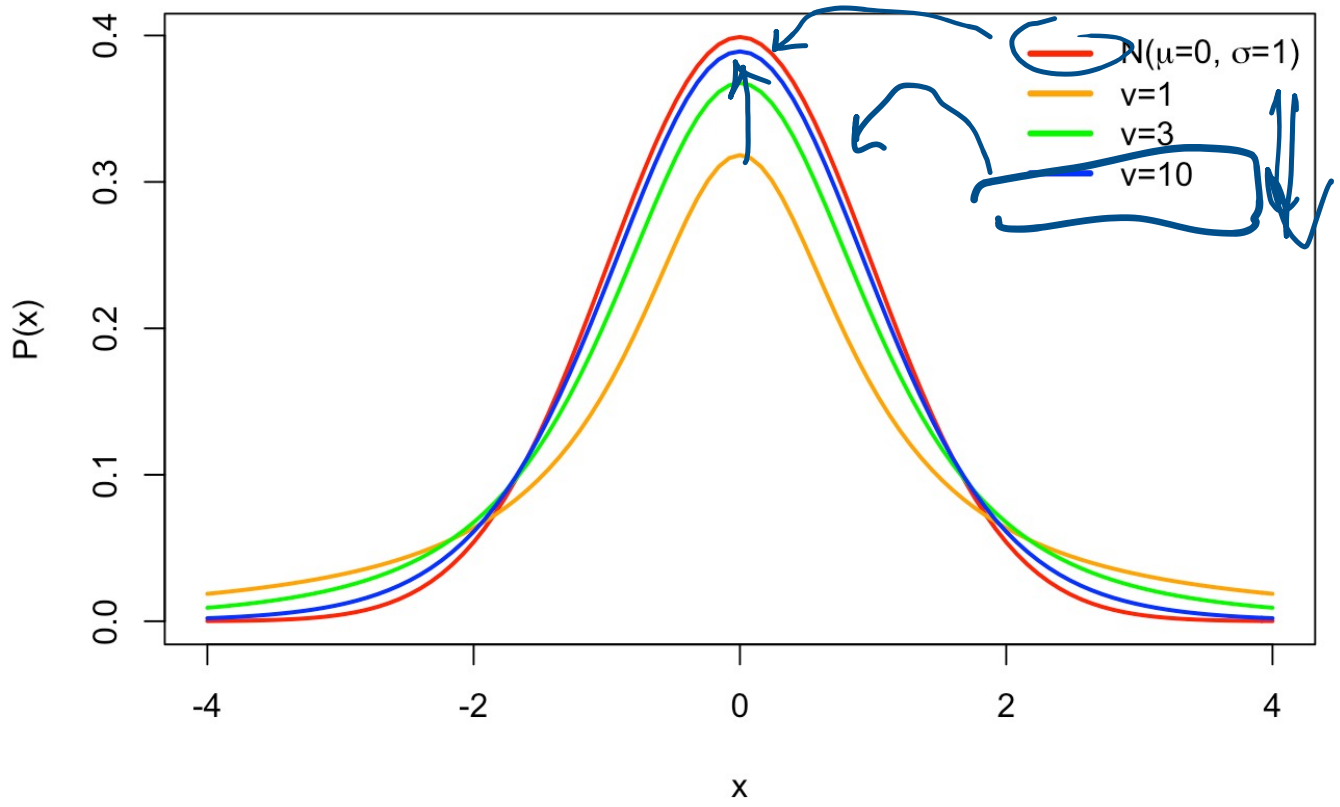
no idea

$$\overline{X} \pm t_{\alpha/2}s/\sqrt{n},$$

std dev

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t \text{ distributed}$$

Why t distribution ?



d.f = n-1

points -1

How to test for differences between samples:

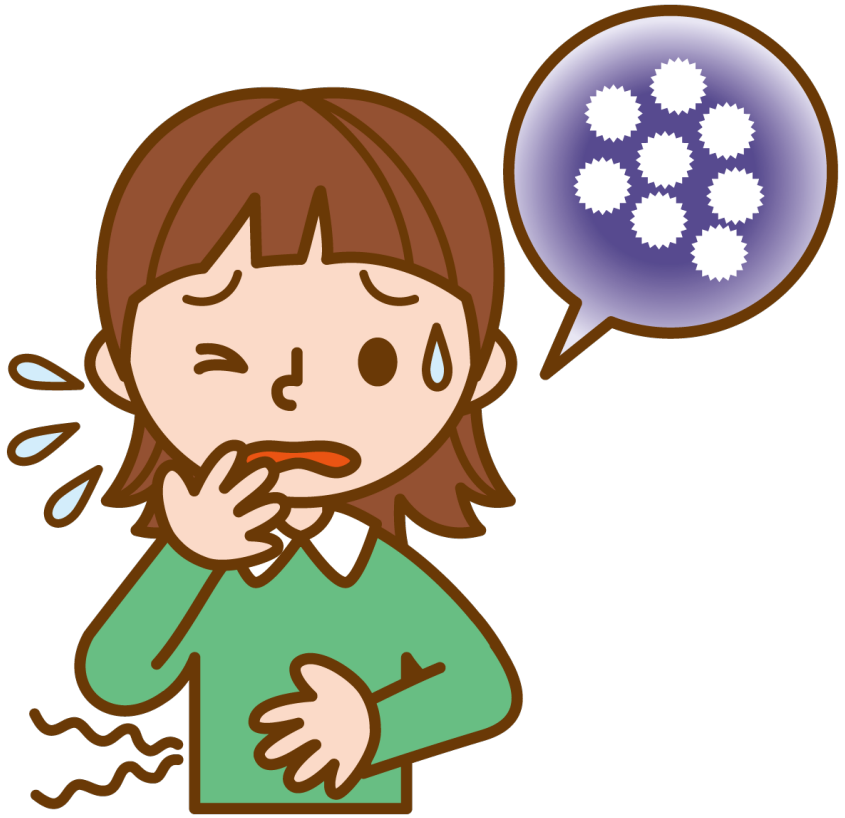
Hypothesis testing

How to test for differences between samples:

Hypothesis testing

Healthy

VS.



Sick

How to test for differences between samples:

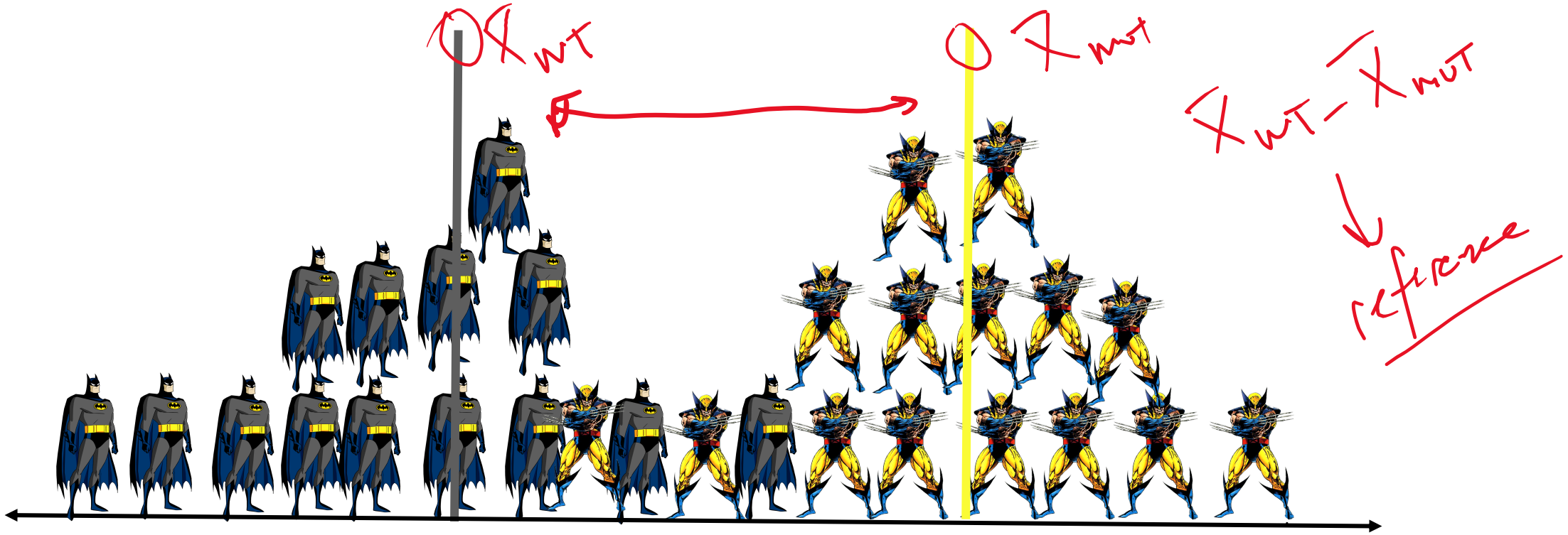
Hypothesis testing

Wild-type (WT)

Mutant (Mut)

How to test for differences between samples:

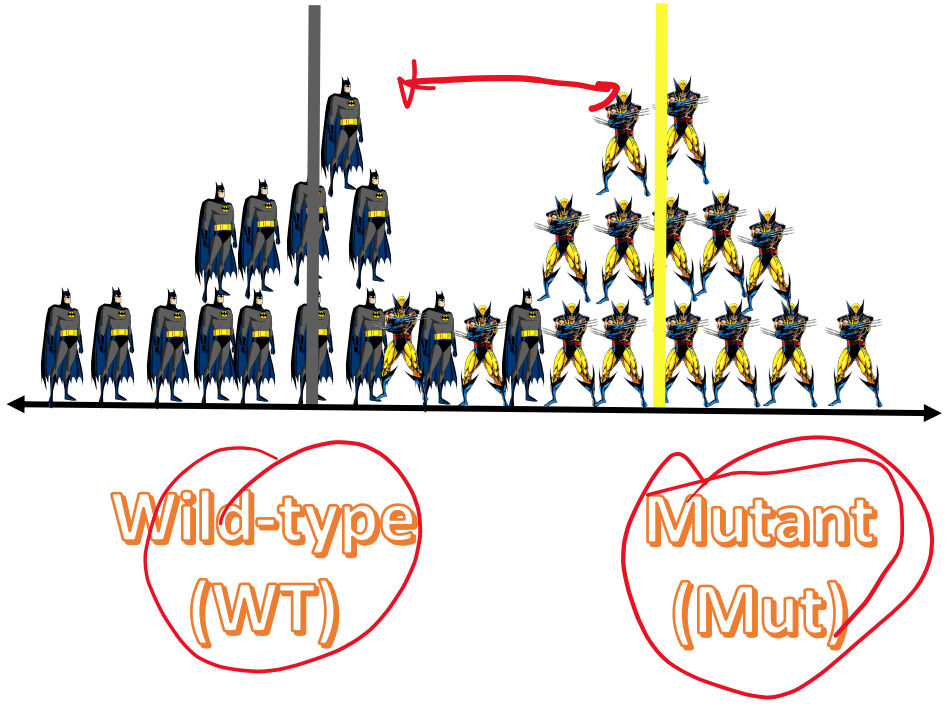
Hypothesis testing: Difference between the means



Wild-type (WT) **Mutant (Mut)**

How to test for differences between samples:

Hypothesis testing: Difference between the means



Core ideas:

Decide on a hypothesis to test, often called “null hypothesis” H_0 H_1

Decide on a statistic to test the truth of the null hypothesis.

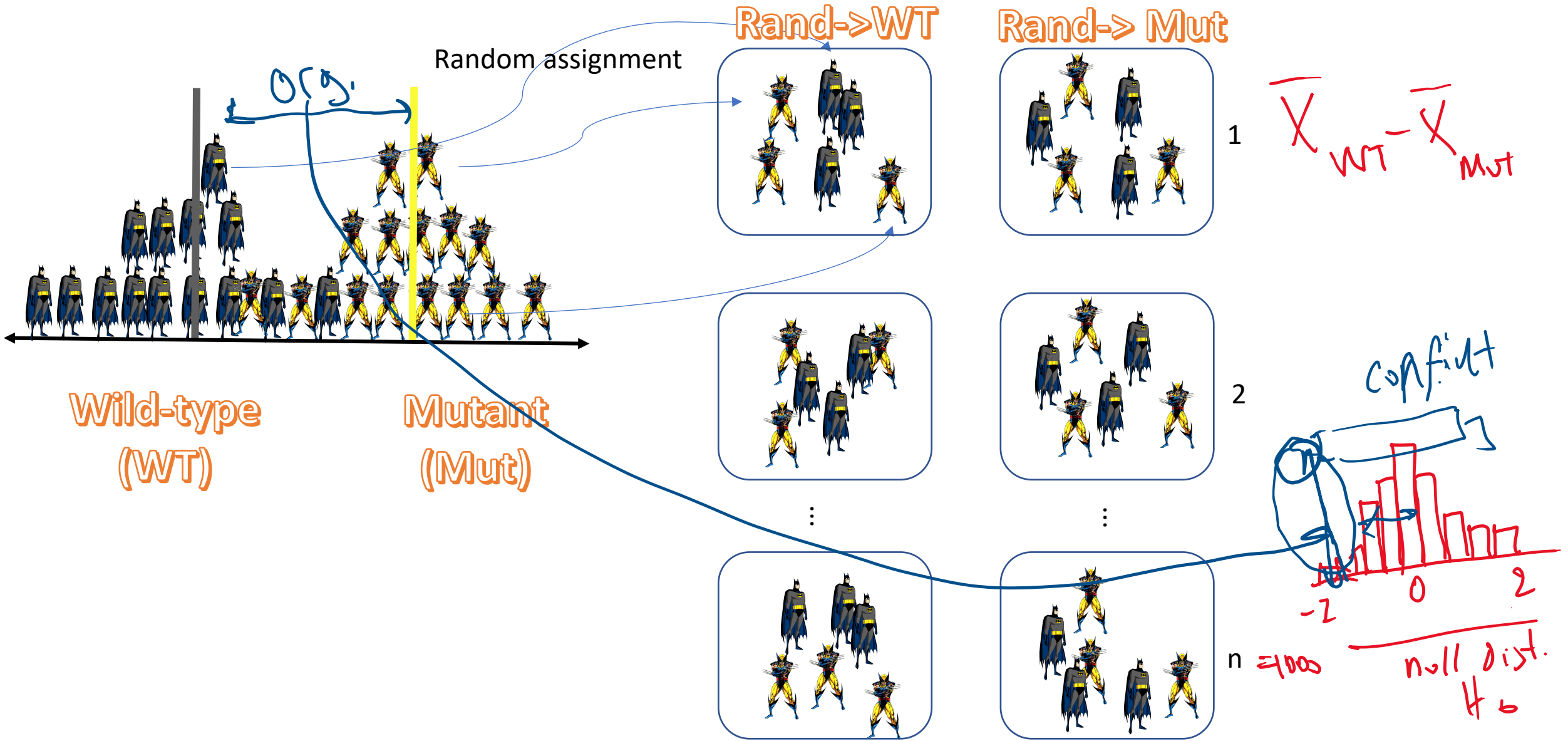
Calculate the statistics and compare to a reference value

$\bar{X}$ diff obs ↔ vs $\bar{X}$ diff ref

H_0

How to test for differences between samples:

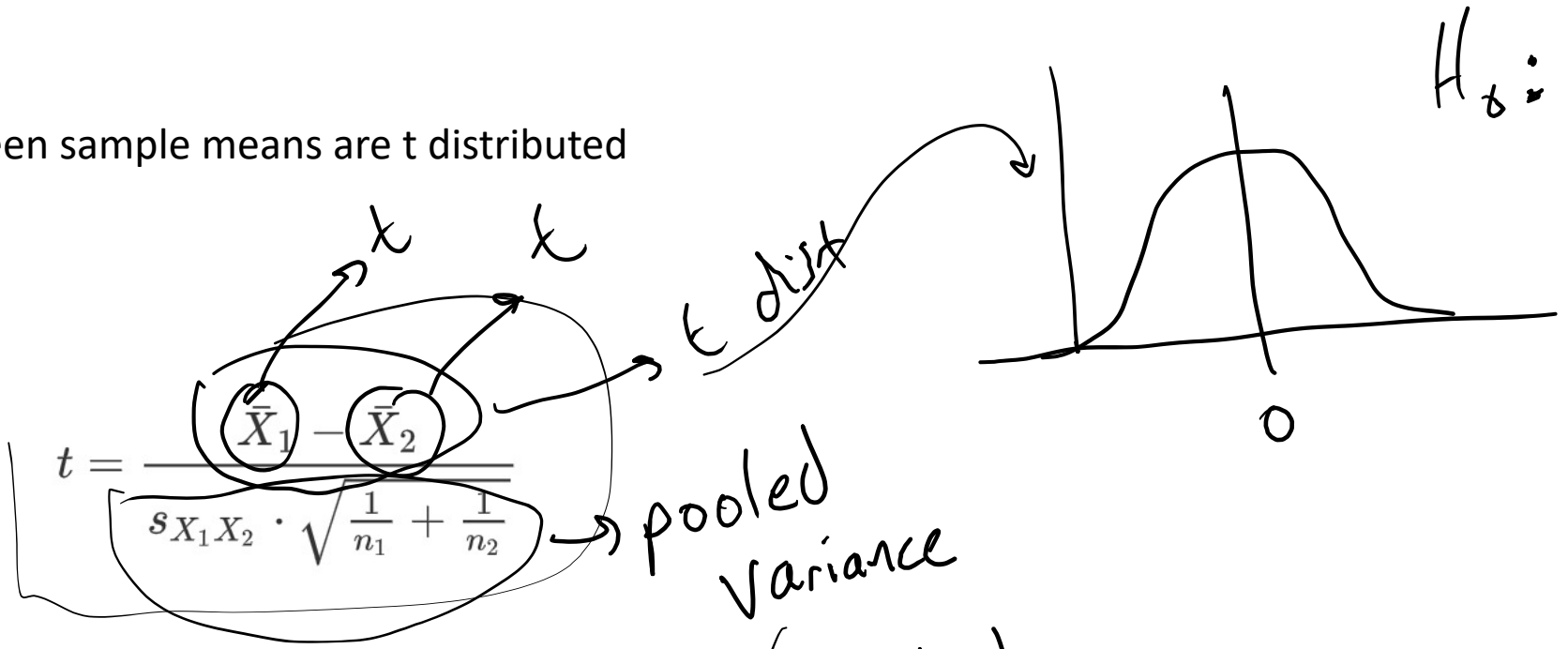
Randomization based test



How to test for differences between samples:

Hypothesis testing: t-test

The differences between sample means are t distributed



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

pooled variance

weighted average

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

→ assume eq. variance

How to test for differences between samples:

Hypothesis testing: t-test

Unequal variance assumption: Welch's t-test

default & t-test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

How to test for differences between samples:

Hypothesis testing: multiple testing correction

	H_0 is TRUE, [Gene is NOT differentially expressed]	H_1 is TRUE, [Gene is differentially expressed]	
Accept H_0 (claim that the gene is not differentially expressed)	True Negatives (TN)	False Negatives (FN) ,type II error	m_0: number of truly null hypotheses
reject H_0 (claim that the gene is differentially expressed)	False Positives (FP) ,type I error	True Positives (TP)	$m - m_0$: number of truly alternative hypotheses

sensitivity

How to test for differences between samples:

Hypothesis testing: multiple testing correction

General idea:

- more FPs as we test more
- Push P-values to higher values to avoid making these mistakes

Bonferroni

$$m \cdot p_i$$

tests

False Discovery Rate

$$\frac{m \cdot p_i}{i}$$

rank

Q values

Depends on estimated proportion of True null hypothesis

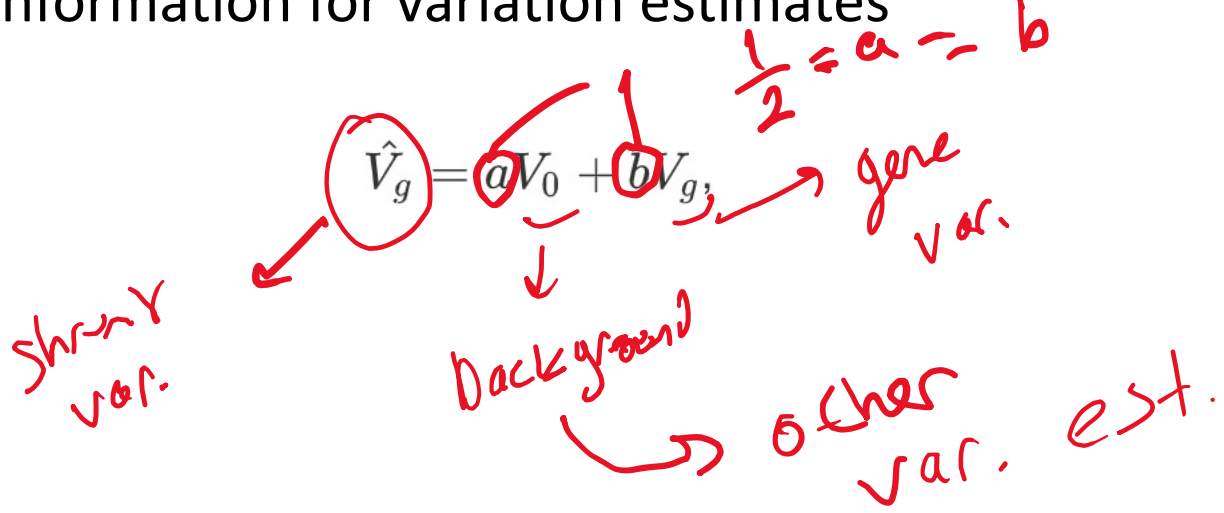
How to test for differences between samples:

Hypothesis testing: moderated t-tests

General idea: Use information from the many tests you are doing

Doing many tests is typical procedure for gene expression analysis

Borrow information for variation estimates



How to test for differences between samples:

Hypothesis testing: moderated t-tests

Moderated tests in action: a simple example

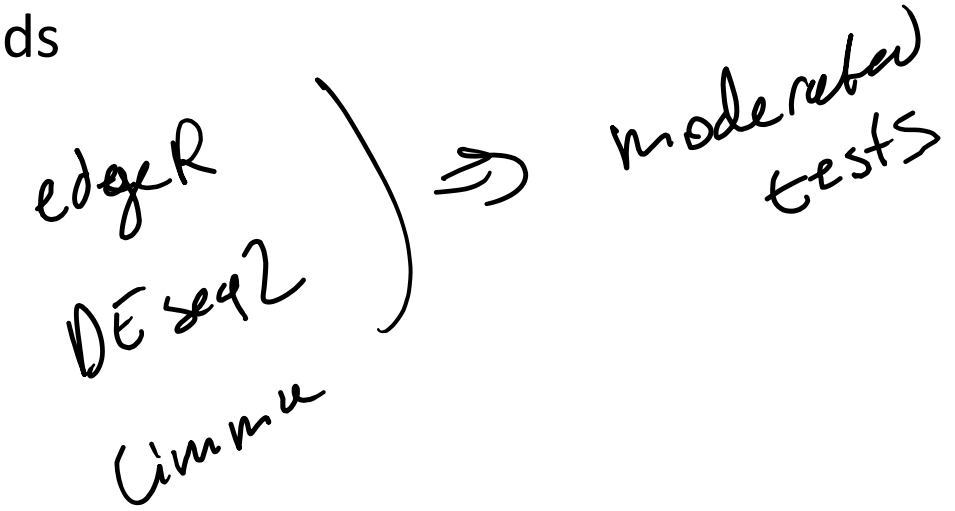
- Simulate 1000 genes, 3 test and 3 control groups
- No real difference in any of the genes
- Observe the effect of moderated and unmoderated tests

$$\hat{V}_g = (V_0 + V_g)/2$$

How to test for differences between samples:

Hypothesis testing: moderated t-tests

Why is this important? moderation of parameter estimates are used by modern diff. exp. methods



Relationship between variables:

Linear regression

We often need to model the relationship between variables

$$Y = \beta_0 + \beta_1 X + \epsilon$$

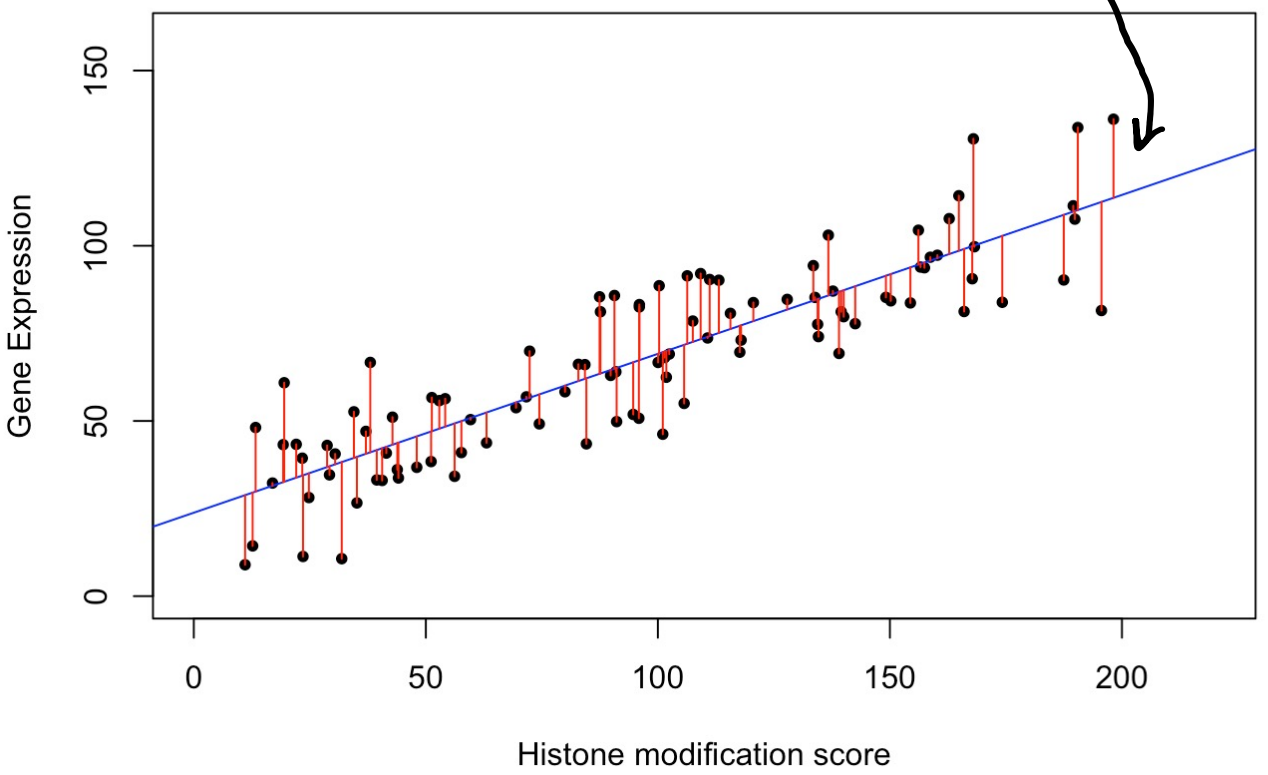
Dependent variable
Response variable

Independent variable
Explanatory variables
Predictor variable

Relationship between variables:

Linear regression

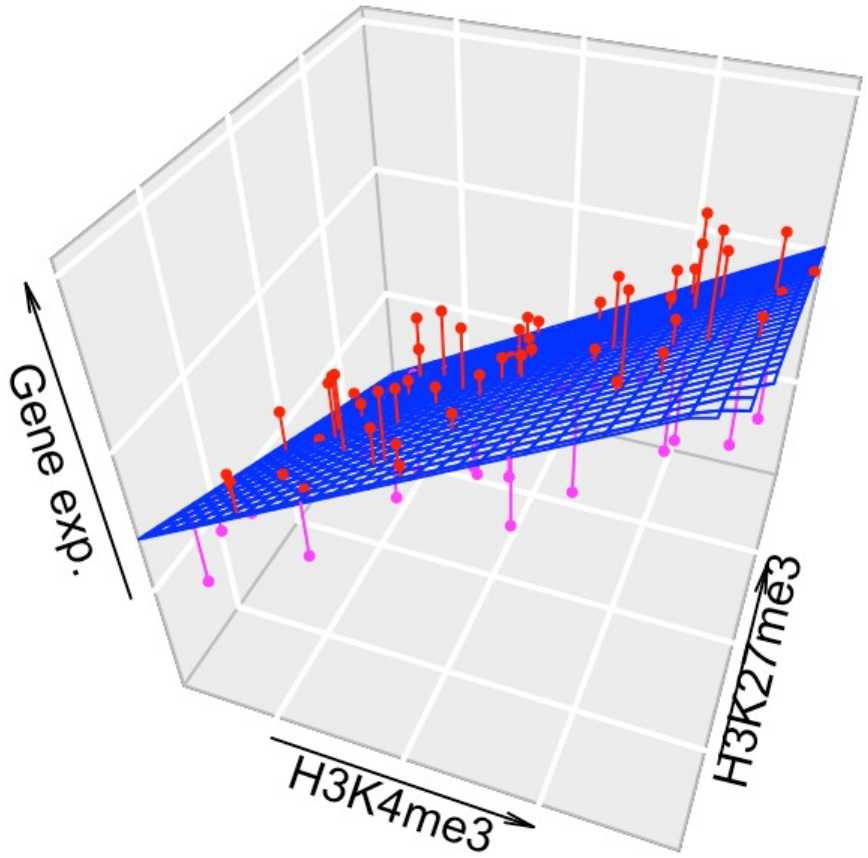
$$Y = \beta_0 + \beta_1 X + \epsilon$$



Relationship between variables:

Linear regression

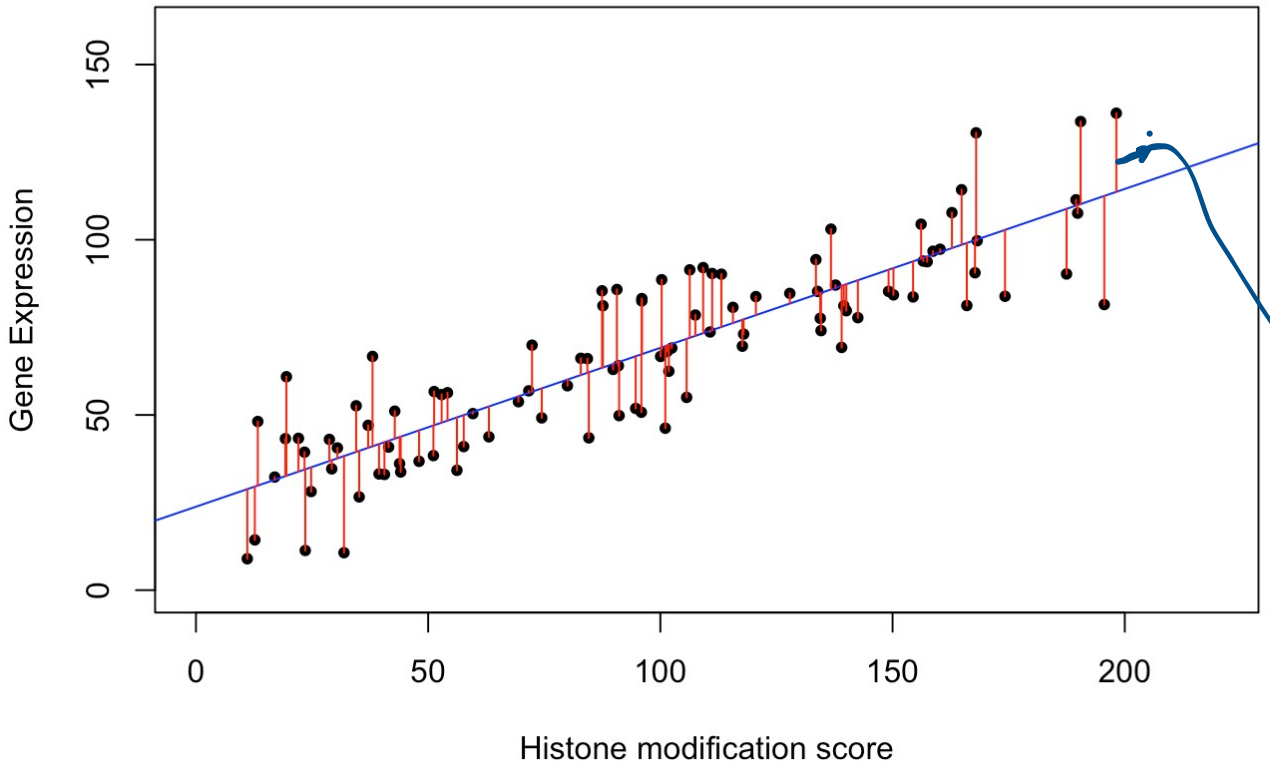
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Relationship between variables:

How to fit a line ?

$$Y = \beta_0 + \beta_1 X + \epsilon$$

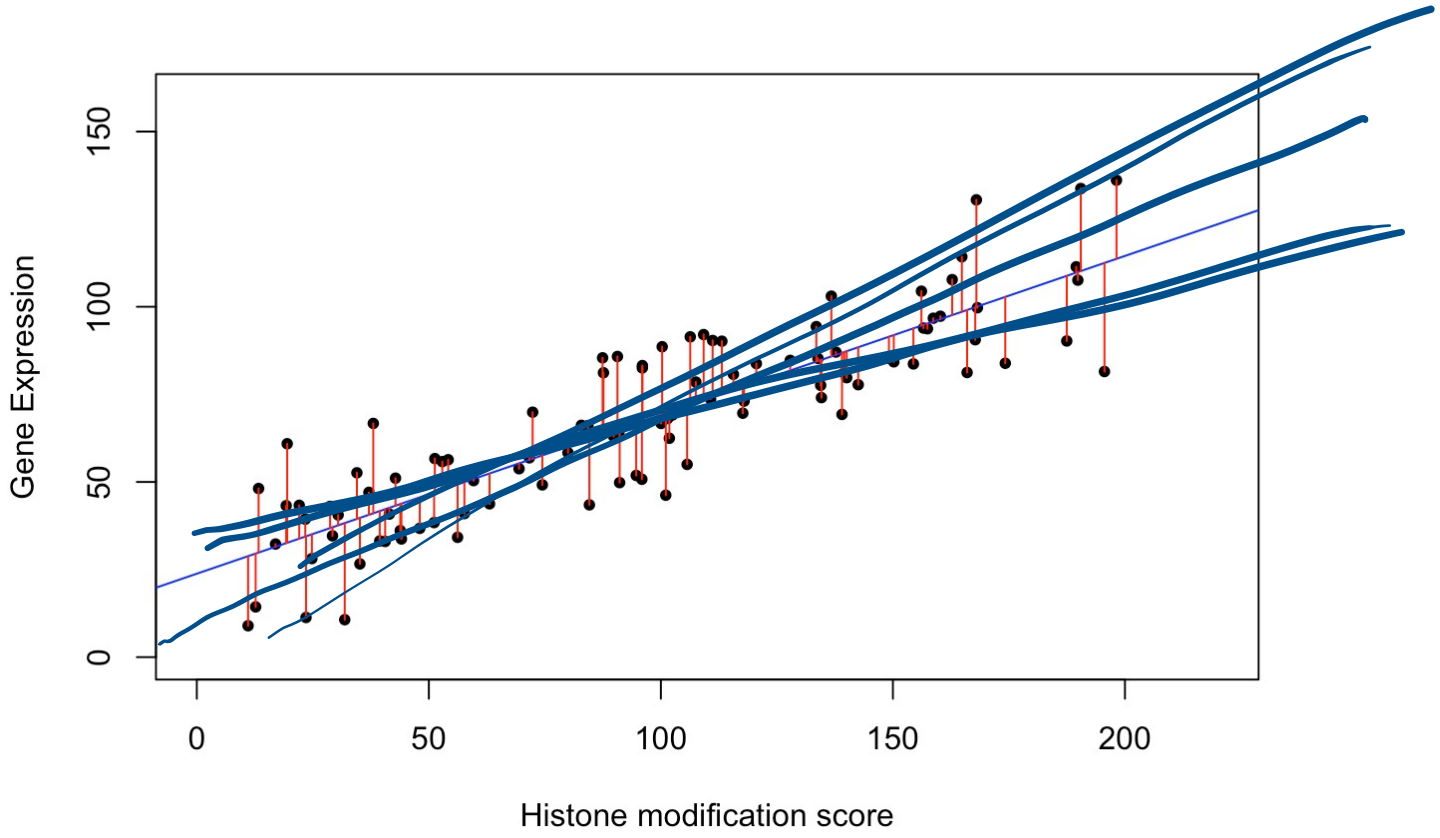


$$min \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

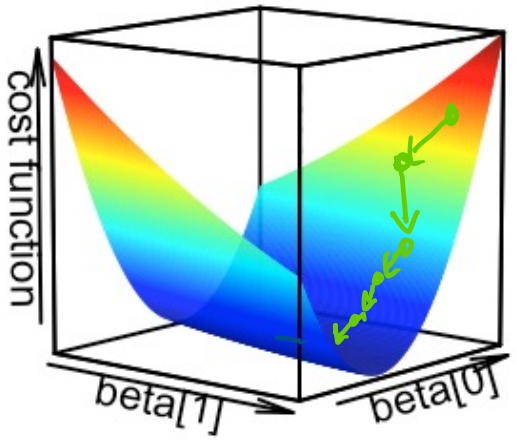
Relationship between variables:

How to fit a line ? -> cost or loss function

$$Y = \beta_0 + \beta_1 X + \epsilon$$



$$min \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

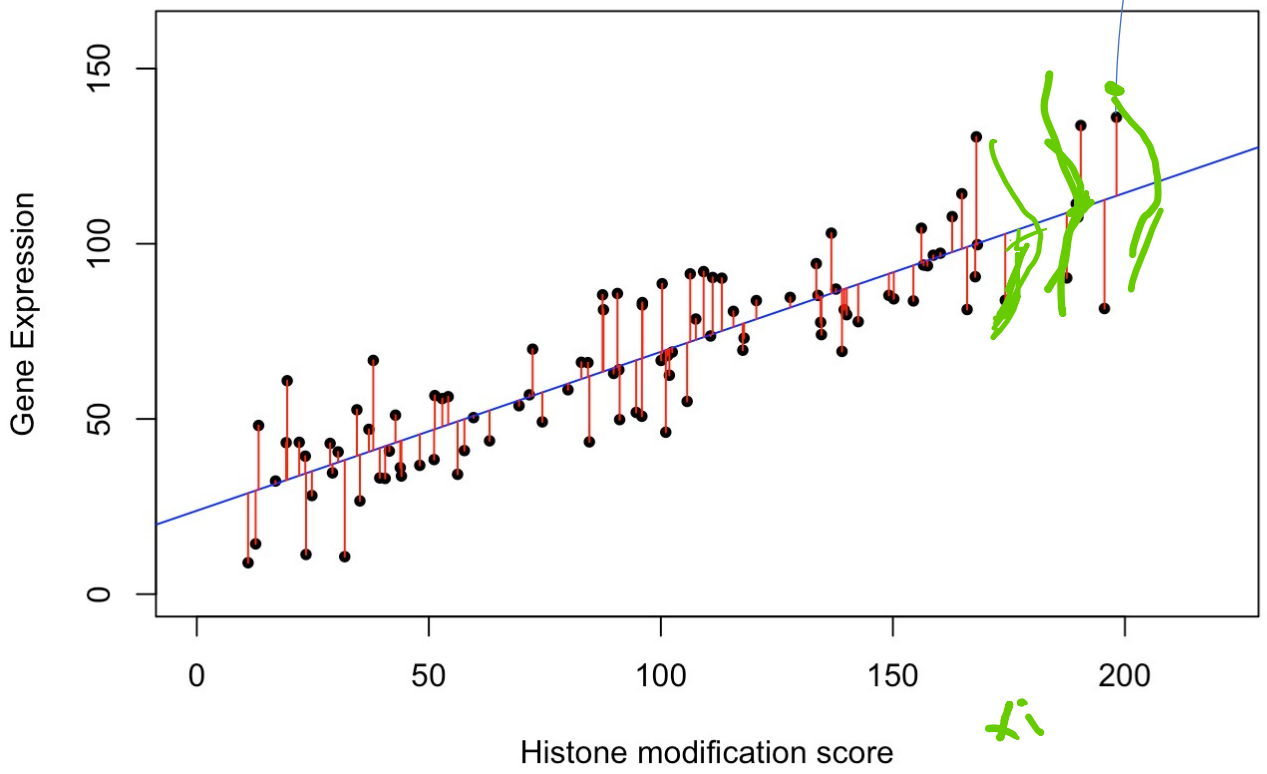


Relationship between variables:

How to fit a line ? -> maximum likelihood function

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$P(y_i) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{s}\right)^2}$$



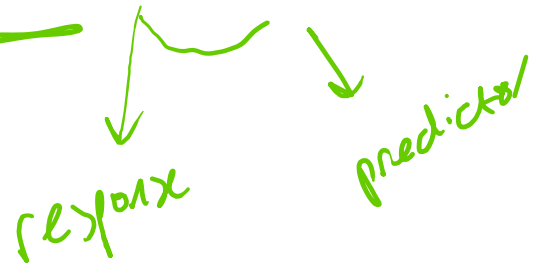
$$L = P(y_1)P(y_2)P(y_3)..P(y_n) = \prod_{i=1}^{n} P_i$$

$$ln(L) = -nln(s\sqrt{2\pi}) - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Relationship between variables:

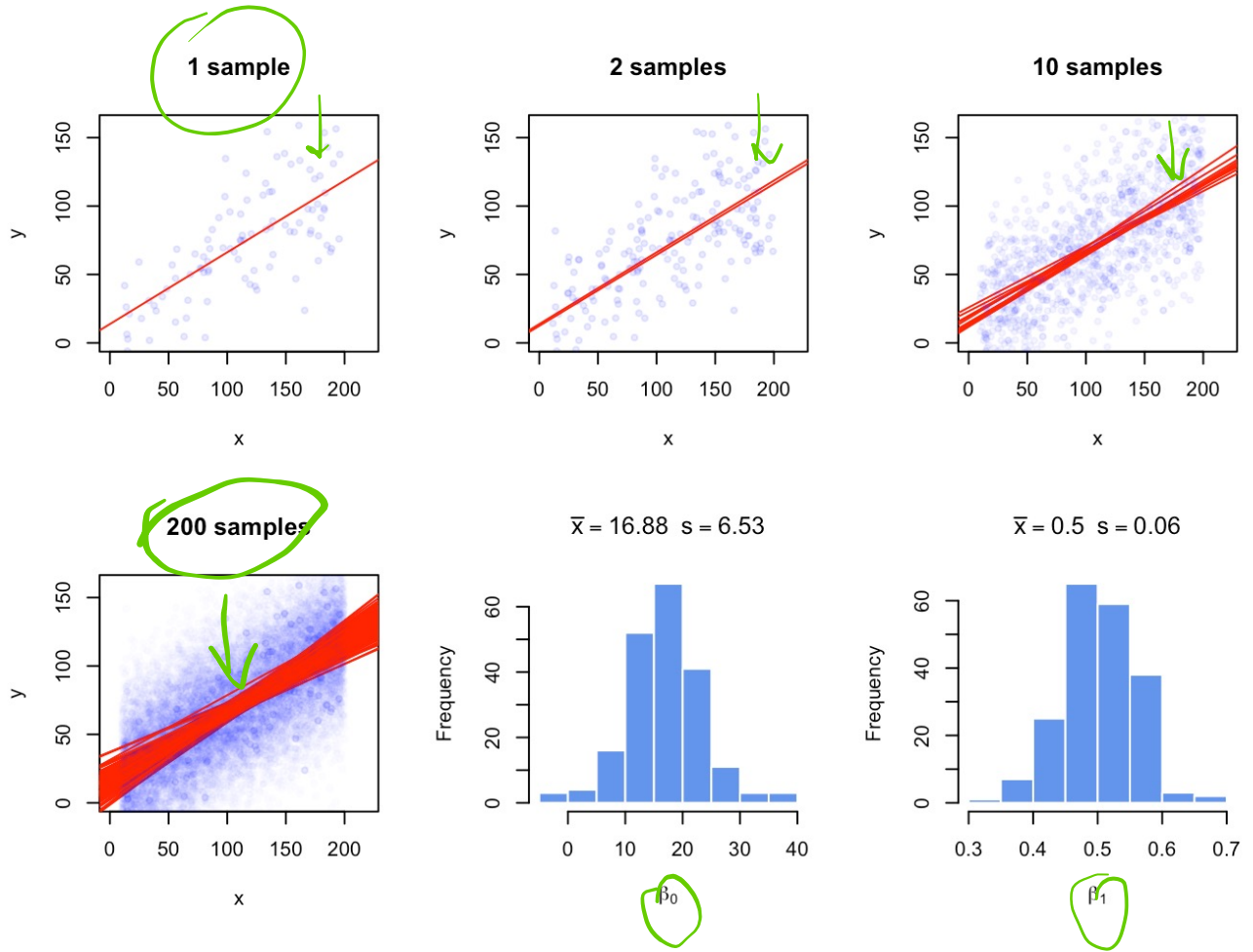
Fitting lines in R

```
mod1=lm(y~x)
```



Relationship between variables:

Error of the coefficients



Relationship between variables:

Error of the coefficients

$$s = RSE = \sqrt{\frac{\sum (y_i - (\beta_0 + \beta_1 x_i))^2}{n-2}} = \sqrt{\frac{\sum \epsilon^2}{n-2}}$$

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \overline{X})^2}}$$

Relationship between variables:

Accuracy of the model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.40951    6.43208   14.06   <2e-16 ***
x            2.08742    0.09775   21.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.96 on 48 degrees of freedom
Multiple R-squared:  0.9048,    Adjusted R-squared:  0.9028
F-statistic:   456 on 1 and 48 DF,  p-value: < 2.2e-16
```

$$s = RSE = \sqrt{\frac{\sum (y_i - \hat{Y}_i)^2}{n - p}} = \sqrt{\frac{RSS}{n - p}}$$

Relationship between variables:

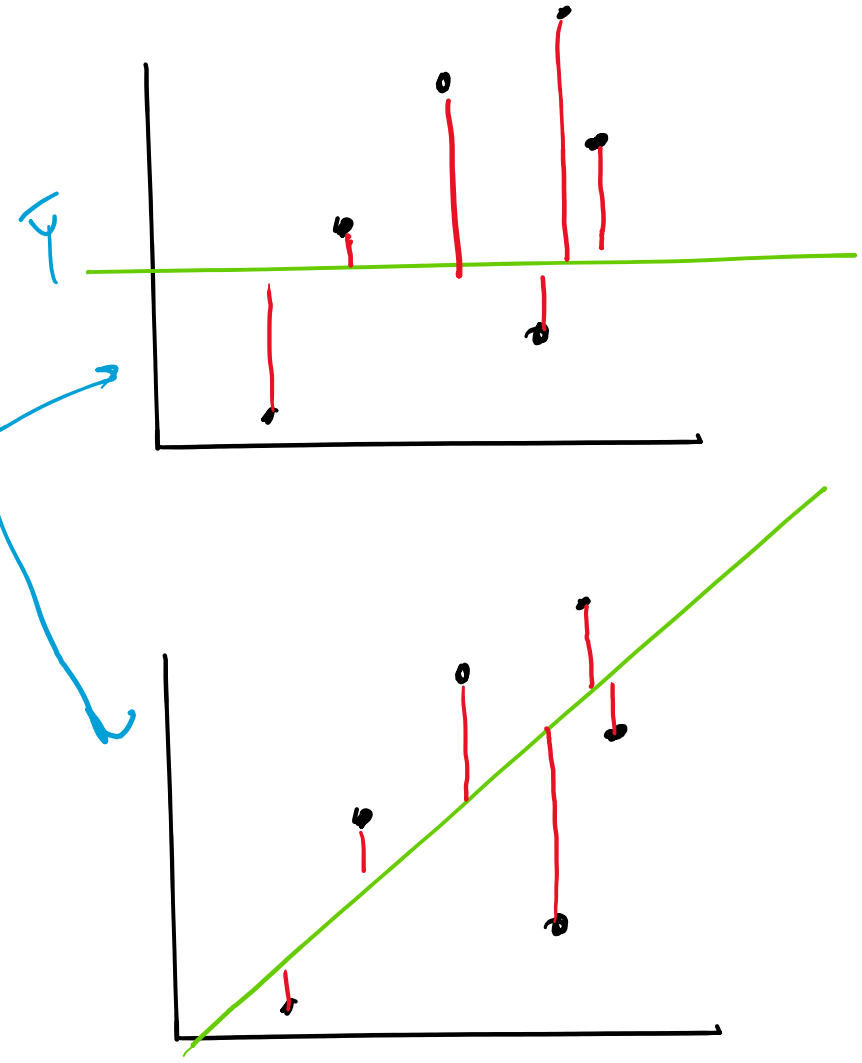
Accuracy of the model

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\sum (y_i - \hat{Y}_i)^2$$

Explained variability : R-squared

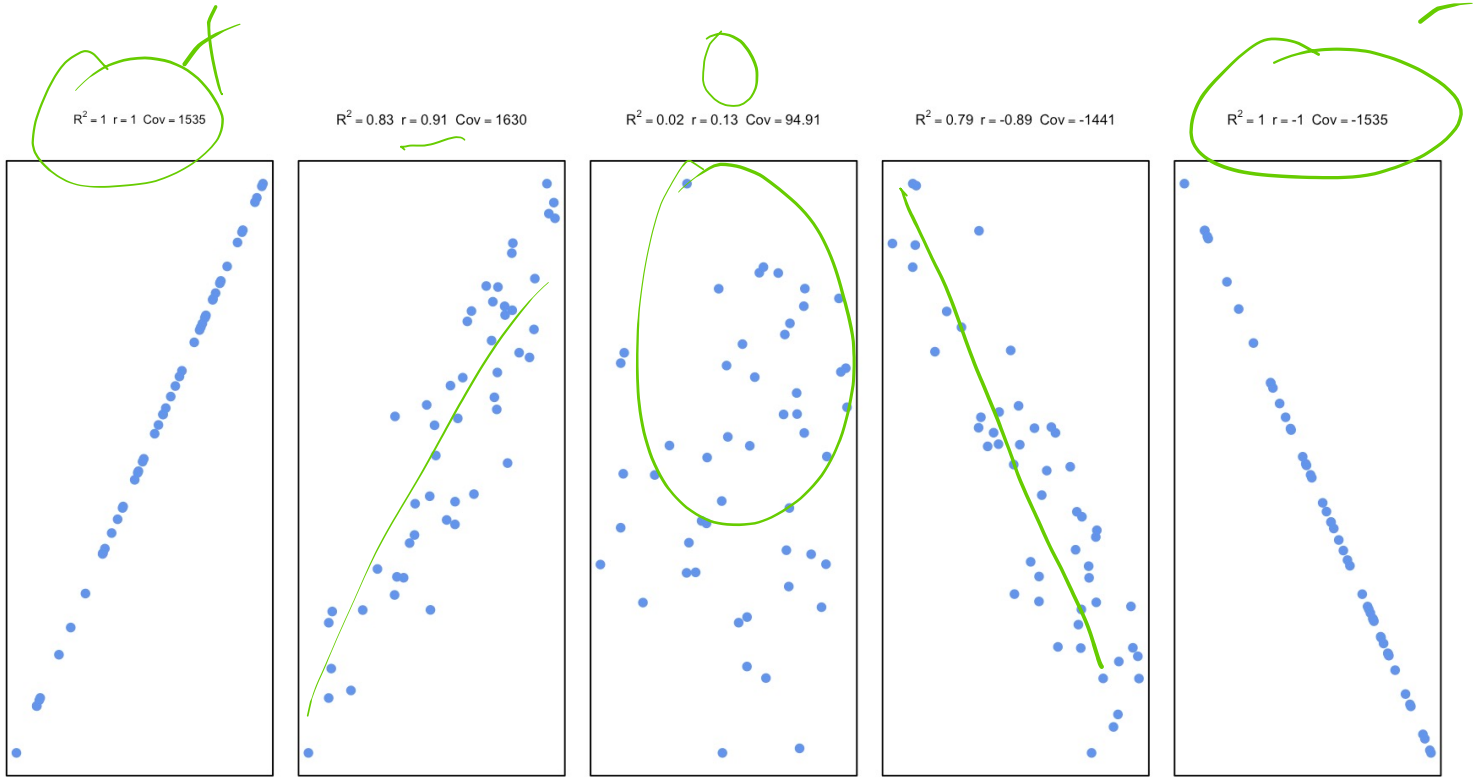
F-statistic → p-value



Relationship between variables:

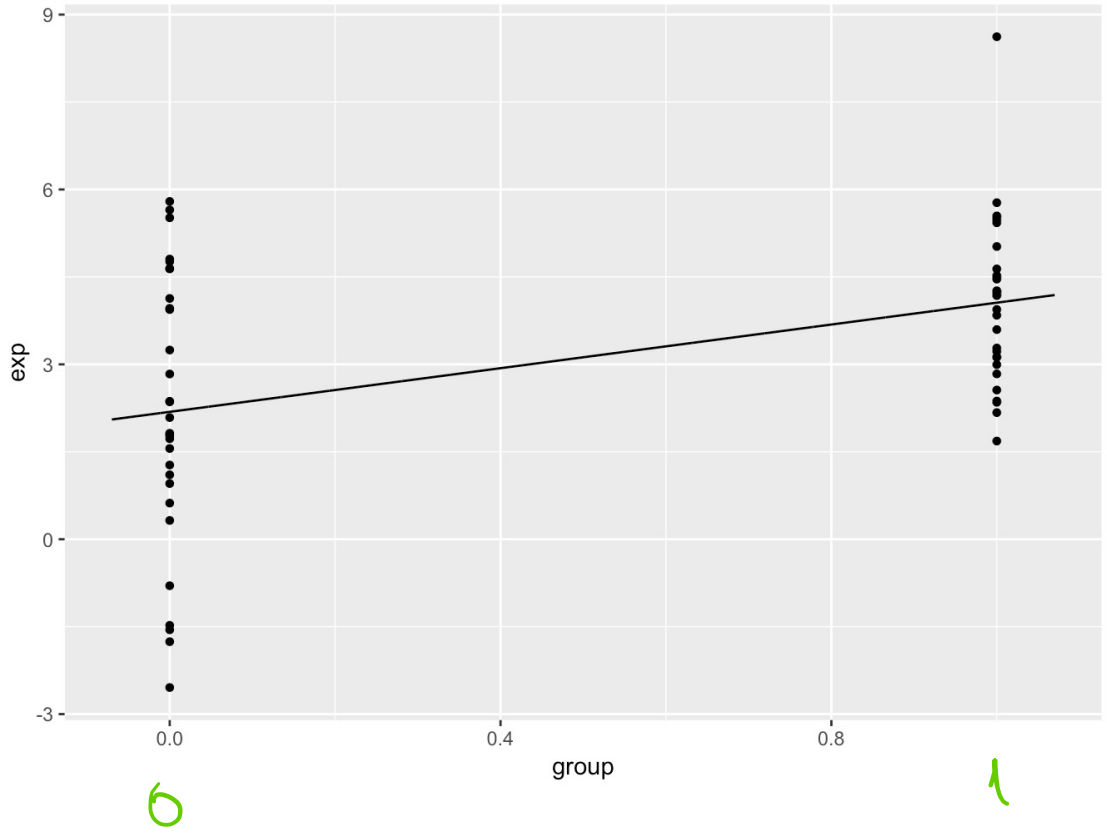
Correlation

$$r_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Relationship between variables:

Regression with categorical variables



Why regression framework for GE?

- Control covariates
- Have more groups

Statistics for Genomics:

Recap

Key concepts:

1) Hypothesis testing
   - Statistical distributions
   - Multiple testing correction
   - Confidence intervals
   - Moderated tests
   - Resampling techniques
2) Linear regression
   - Optimization
   - Model accuracy

ML