A statistics refresher

Karl W Broman

Department of Biostatistics
Johns Hopkins University

kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

Outline

- What is statistics?
- Populations, samples
- Parameters, estimates
- Sampling distributions
- Bias, SE, MSE
- Maximum likelihood
- Confidence intervals
- Hypothesis tests
- Multiple-sample problem
- Linear regression
- Bayesian statistics

Caveat: We're covering a lot of ground in a short period of time, and so lots of approximations will be made.

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir R. A. Fisher

What is statistics?

- Data exploration and analysis
- Inductive inference with probability
- Quantification of uncertainty
- Experimental design

Populations and samples

We are interested in the distribution of some measurement (or measurements) in some underlying (possibly hypothetical) population (or populations).

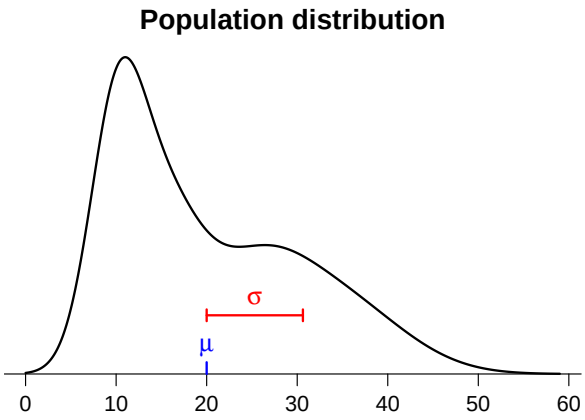
Examples:

- Infinite number of mice from strain A; cytokine response to treatment.
- All T cells in a person; respond or not to an antigen.
- All possible samples from the Baltimore water supply; concentration of cryptospiridium.
- All possible samples of a particular type of cancer tissue; expression of a certain gene.

We can't see the entire population (whether it is real or hypothetical), but we can see a random sample of the population (perhaps a set of independent, replicated measurements).

Parameters

The object of our interest is the population distribution or, in particular, certain numerical attributes of the population distribution (called parameters).



Examples:

- mean
- median
- SD
- proportion = 1
- proportion $>$ 40
- geometric mean
- 95th percentile

Parameters are usually assigned greek letters (like θ, μ, and σ).

Sample data

We make n independent measurements (or draw a random sample of size n).

This gives $X_1, X_2, \ldots, X_n$, independent and identically distributed (iid), following the population distribution.

Statistic: A numerical summary (function) of the X's (that is, of the data). For example, the sample mean, sample SD, etc.

Estimator: (estimate) A statistic, viewed as estimating some population parameter.

We write:

$\hat{\theta}$ an estimator of θ

$\hat{p}$ an estimator of p

$\bar{X} = \hat{\mu}$ an estimator of μ

$S = \hat{\sigma}$ an estimator of σ

Parameters, estimators, estimates

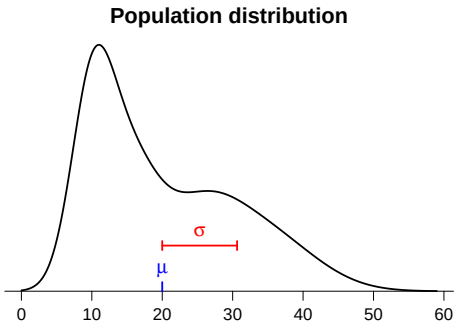
μ
- The population mean
- A parameter
- A fixed quantity
- Unknown, but what we want to know

$\bar{X}$
- The sample mean
- An estimator of μ
- A function of the data (the X's)
- A random quantity

$\bar{x}$
- The observed sample mean
- An estimate of μ
- A particular realization of the estimator, $\bar{X}$
- A fixed quantity, but the result of a random process.

Estimators are random variables

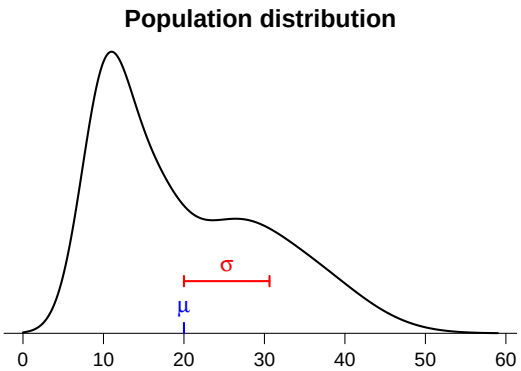
Estimators have distributions, means, SDs, etc.



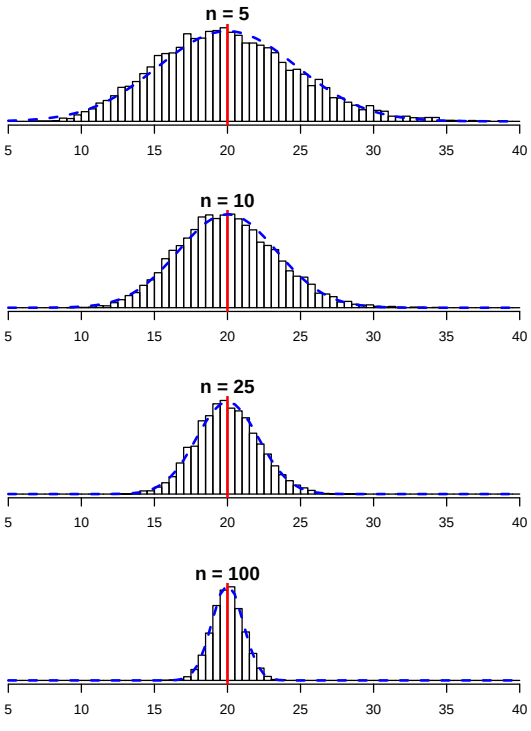
$\longrightarrow \quad X_1, X_2, \ldots, X_{10} \quad \longrightarrow \quad \bar{X}$

3.8	8.0	9.9	13.1	15.5	16.6	22.3	25.4	31.0	40.0	$\longrightarrow$ 18.6
6.0	10.6	13.8	17.1	20.2	22.5	22.9	28.6	33.1	36.7	$\longrightarrow$ 21.2
8.1	9.0	9.5	12.2	13.3	20.5	20.8	30.3	31.6	34.6	$\longrightarrow$ 19.0
4.2	10.3	11.0	13.9	16.5	18.2	18.9	20.4	28.4	34.4	$\longrightarrow$ 17.6
8.4	15.2	17.1	17.2	21.2	23.0	26.7	28.2	32.8	38.0	$\longrightarrow$ 22.8

Sampling distribution



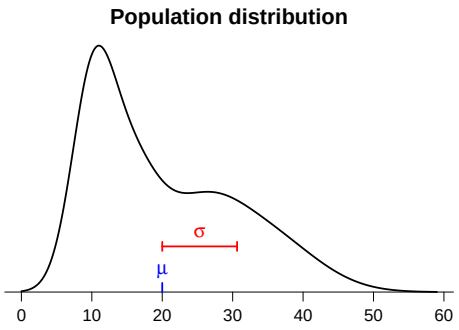
Distribution of $\bar{X}$

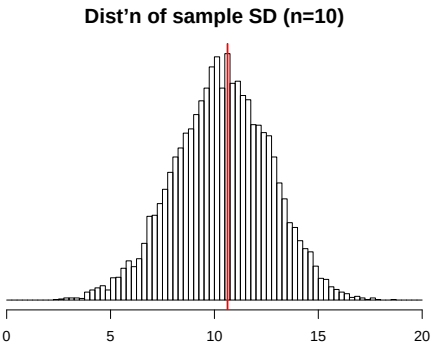


Sampling distribution depends on:

- The type of statistic
- The population distribution
- The sample size

Bias, SE, RMSE





Consider $\hat{\theta}$, an estimator of the parameter θ.

Bias:	$\mathsf{E}(\hat{\theta} - \theta) = \mathsf{E}(\hat{\theta}) - \theta$
Standard error (SE):	$\mathsf{SE}(\hat{\theta}) = \mathsf{SD}(\hat{\theta})$.
RMS error (RMSE):	$\sqrt{\mathsf{E}\{(\hat{\theta} - \theta)^2\}} = \sqrt{(\mathsf{bias})^2 + (\mathsf{SE})^2}$.

An example

Consider n backcross mice (genotype AA or AB). Let y_i be the quantitative phenotype of mouse i.

Imagine a single gene (QTL) + environmental noise.

Let $z_i = 1/0$ if mouse i is AB/AA at the gene, and suppose there is no segregation distortion, so that $\Pr(z_i = 0) = \Pr(z_i = 1) = 1/2$.

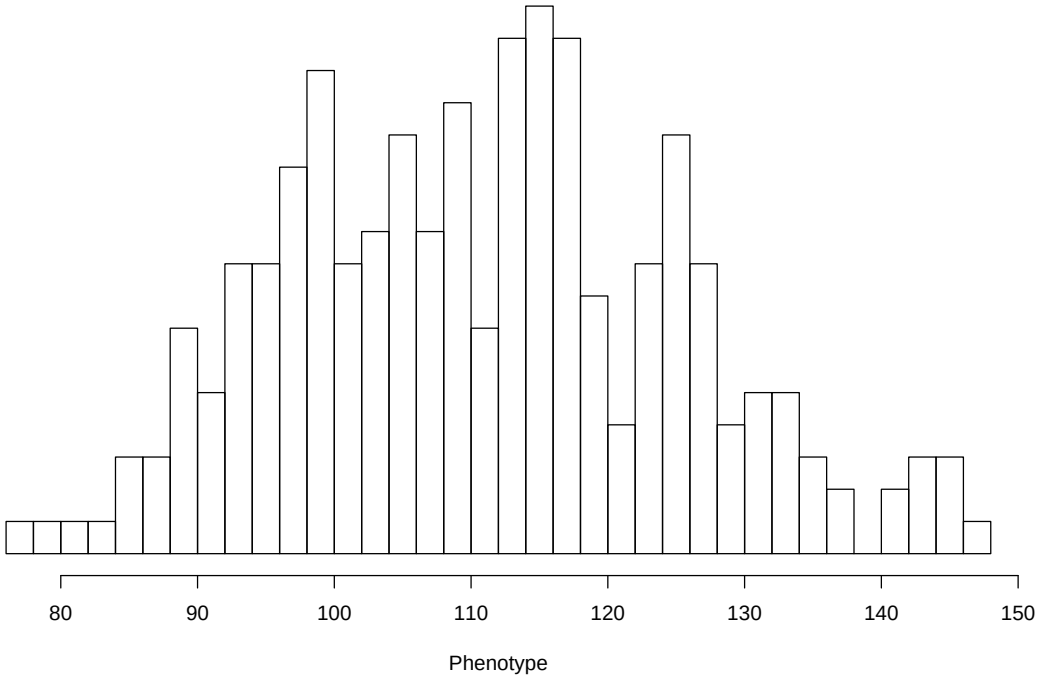
Imagine $y_i|z_i \sim \text{normal}(\mu_{z_i}, \sigma^2)$.

$$f(y_i|z_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu_{z_i}}{\sigma}\right)^2\right]$$

Unconditionally, y_i follows a mixture of normal distributions.

$$f(y_i) = (1/2)\ f(y_i|z_i = 1) + (1/2)\ f(y_i|z_i = 0)$$

Example data



Maximum likelihood

Q: How to estimate μ_0, μ_1, and σ?

One approach: Use the parameter values for which the data are most probable.

log likelihood:

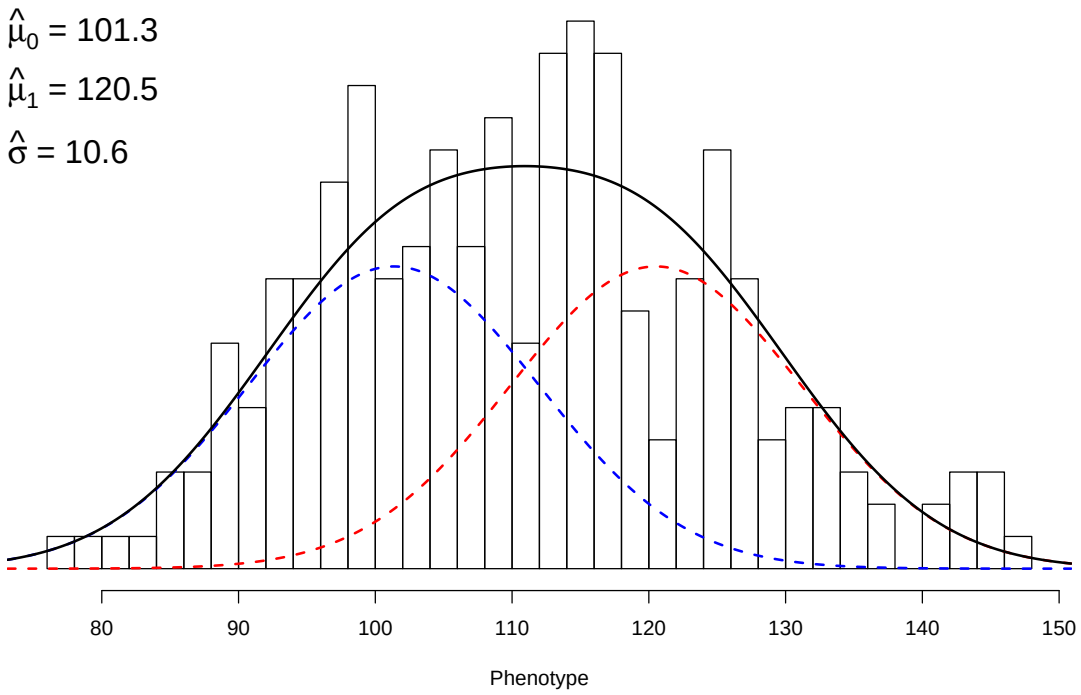
$$\begin{aligned} l(\mu_0, \mu_1, \sigma) &= \log \Pr(\text{data}|\mu_0, \mu_1, \sigma) \\ &= \sum_i \log\left[(1/2)\phi(y_i|\mu_0, \sigma) + (1/2)\phi(y_i|\mu_1, \sigma)\right] \end{aligned}$$

where $\phi(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\frac{y_i-\mu}{\sigma})^2\right]$

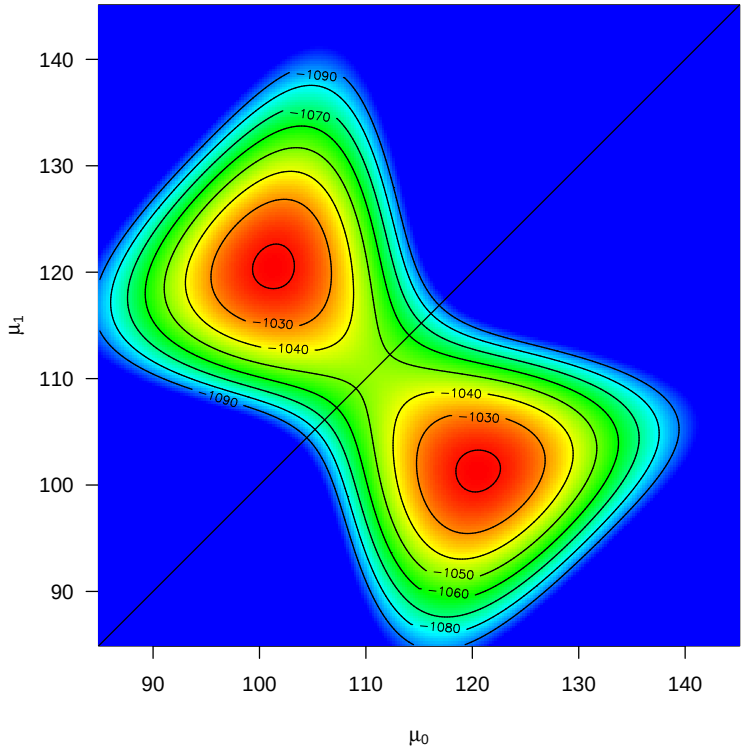
Maximum likelihood estimates:

$\hat{\mu}_0$, $\hat{\mu}_1$, σ such that l is maximized.

Fitted curves



Log likelihood surface



A few facts

Let $X_1, X_2, \ldots, X_n$ be iid with mean = μ and SD = σ.

Let $\bar{X} = \sum X_i/n =$ sample mean

$s = \sqrt{\sum (X_i - \bar{X})^2/(n-1)} =$ sample SD

No matter what: $\mathrm{E}(\bar{X}) = \mu$, $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$.

$\mathrm{E}(s^2) = \sigma^2$

If n is large, $\bar{X}$ is approximately normally distributed

If the population distribution is normal:

$\bar{X} \sim$ normal $\qquad s^2(n-1)/\sigma^2 \sim \chi^2_{n-1}$

$(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathrm{normal}(0,1) \qquad (\bar{X} - \mu)/(s/\sqrt{n}) \sim t_{n-1}$

Confidence intervals

Suppose we measure the $\log_{10}$ cytokine response in 100 male mice of a certain strain, and find that the sample average ($\bar{x}$) is 3.52 and sample SD (s) is 1.62.

Our estimate of the SE of the sample mean is $1.62/\sqrt{100} = 0.162$.

A 95% confidence interval for the population mean (μ) is

$3.52 \pm (2 \times 0.16) = 3.52 \pm 0.32 =$ (3.20, 3.84).

What does this mean?

What is the chance that (3.20, 3.84) contains μ?

What is a confidence interval?

A 95% confidence interval is an interval calculated from the data that in advance has a 95% chance of covering the population parameter.

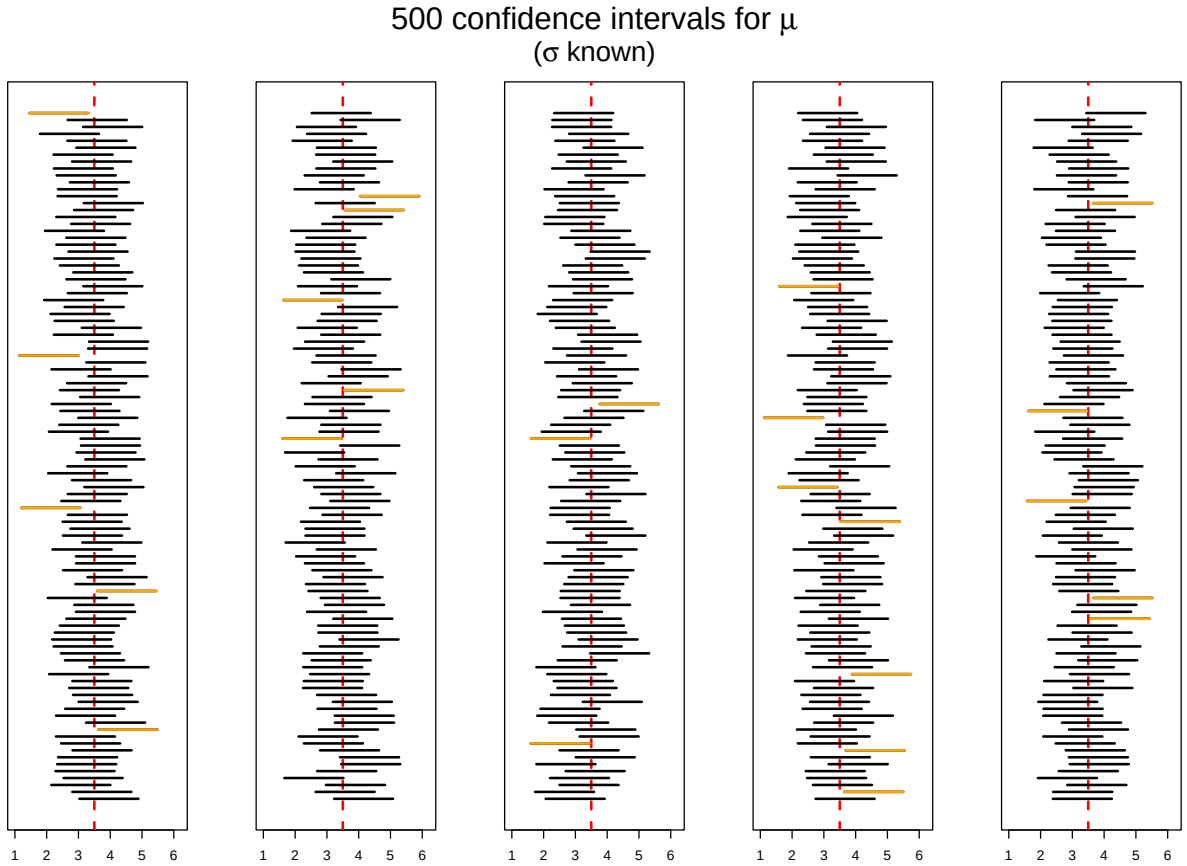
In advance, $\bar{X} \pm 1.96\sigma/\sqrt{n}$ has a 95% chance of covering μ.

Thus, it is called a 95% confidence interval for μ.

Note that, after the data is gathered (for instance, n=100, $\bar{X} = 3.52$, s $= 1.62$), the interval becomes fixed:

$\bar{X} \pm 1.96\sigma/\sqrt{n} = 3.52 \pm 0.32.$

We can't say that there's a 95% chance that μ is in the interval 3.52 ± 0.32. It either is or it isn't; we just don't know.



But we don't know the SD

Use of $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$ as a 95% confidence interval for μ requires knowledge of σ.

That the above is a 95% confidence interval for μ is a result of the following:

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

What if we don't know σ?

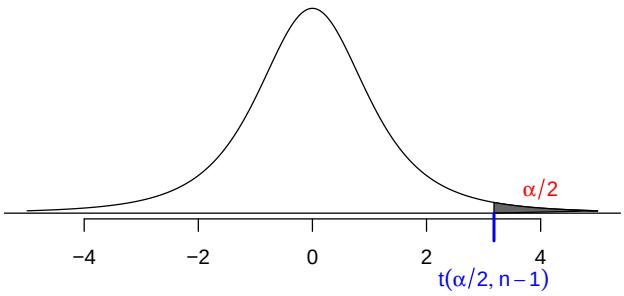
We plug in the sample SD (s), but then we need to widen the intervals to account for the uncertainty in s.

The t interval

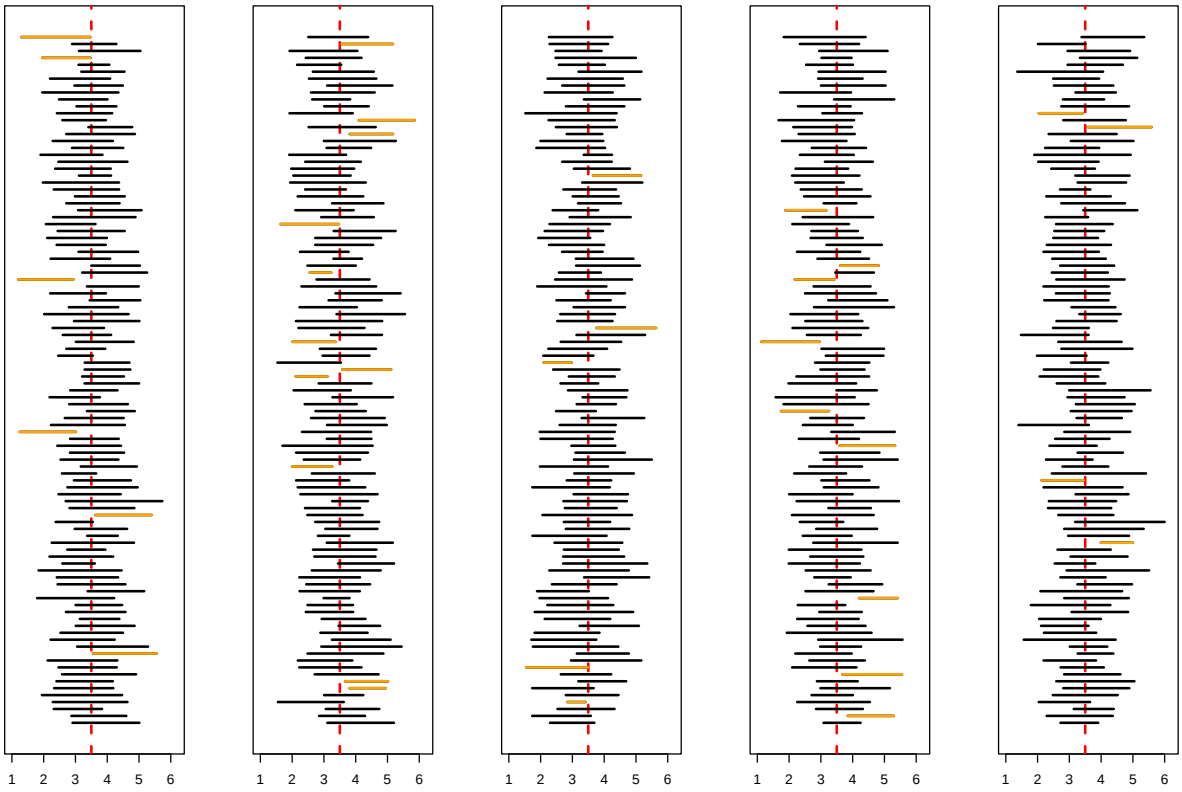
If $X_1, \ldots, X_n$ are iid normal(mean=μ, SD=σ),

$\bar{X} \pm t(\alpha/2, n-1)\; s/\sqrt{n}$ is a $1-\alpha$ confidence interval for μ.

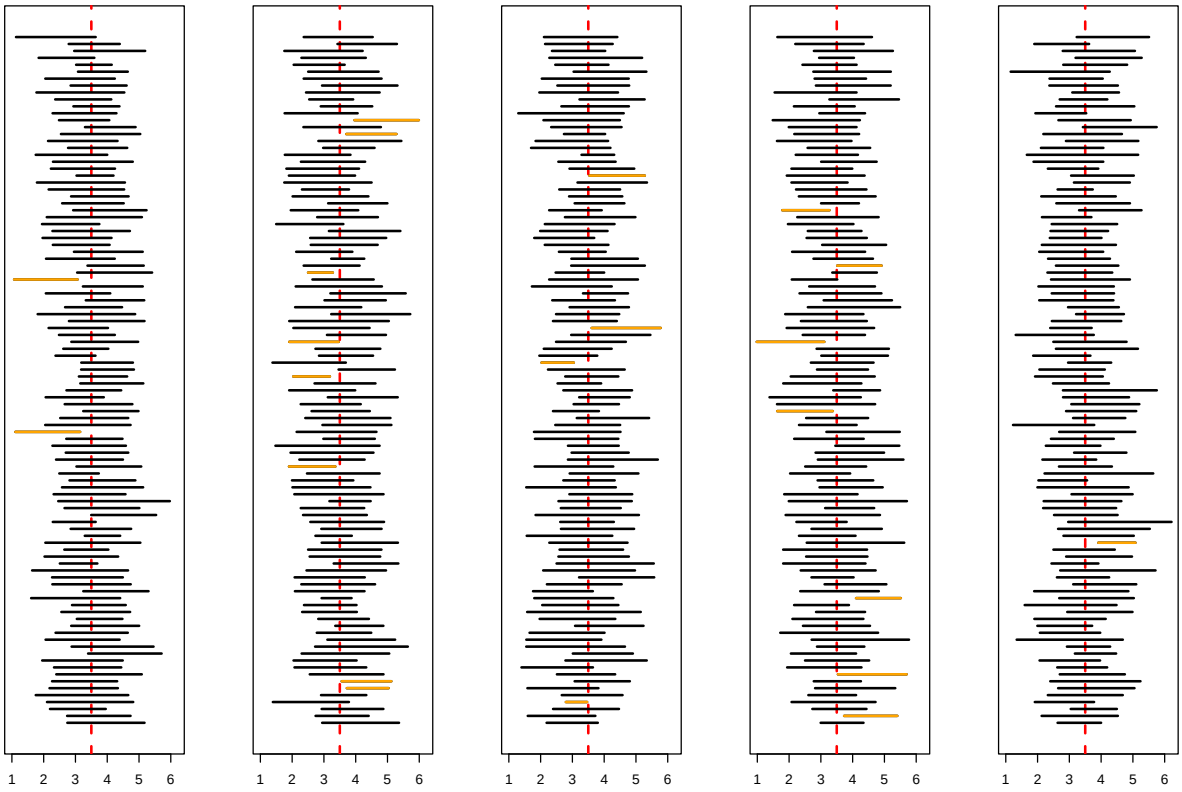
$t(\alpha/2, n-1)$ is the $1-\alpha/2$ quantile of the t distribution with n – 1 "degrees of freedom."



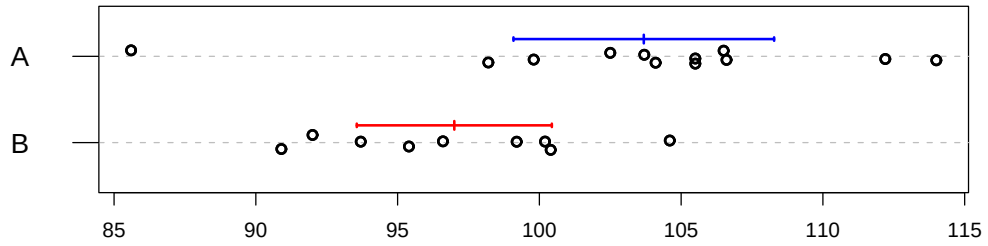
500 BAD confidence intervals for μ
(σ unknown)



500 confidence intervals for μ
(σ unknown)



Hypothesis testing



Question: Do the two strains have the same mean?

We imagine $X_1, \ldots, X_n \sim$ iid normal(μ_A,σ_A)
$Y_1, \ldots, Y_m \sim$ iid normal(μ_B,σ_B)

$H_0 : \mu_A = \mu_B$ $\quad$ $H_a : \mu_A \neq \mu_B$

Question: Are the data compatible with H_0?

Test statistic

In order to determine whether the data are compatible with H_0, we form a summary statistic, for which large values indicate evidence for a departure from the null hypothesis $\mu_A = \mu_B$.

The statistic to use depends on

(a) the types of parameters in question

(b) the form of the data

(c) our assumptions about the process generating the data

In the above example, we'd use T = $\dfrac{\bar{X} - \bar{Y}}{\widehat{\mathrm{SD}}(\bar{X} - \bar{Y})}$

Rejection rule: Reject H_0 if $|T| > C$, for some "critical value," C.

P-values

- P-values are a function of the data. (They are random, like data.)
- P-values measure the strength of evidence against H_0. (Take this with a grain of salt.)
- Small p-values indicate evidence against H_0.
- P = probability of getting this sort of extreme data, if the observed difference were just due to chance variation.
- **NOT** the probability that the observed difference is due to chance.
- Note that P=0.048 is essentially the same as P=0.053.

Permutation test

X or Y	group
X_1	1
X_2	1
⋮	1
X_n	1
Y_1	2
Y_2	2
⋮	⋮
Y_m	2

$\rightarrow T_{obs}$

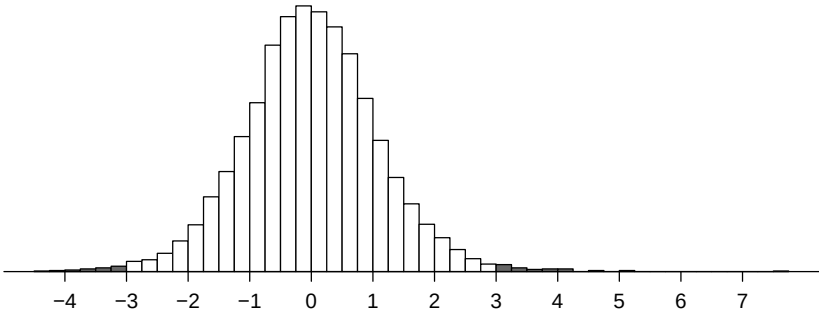
X or Y	group
X_1	2
X_2	2
⋮	1
X_n	2
Y_1	1
Y_2	2
⋮	⋮
Y_m	1

$\rightarrow T^*$

Group status shuffled

Compare the observed t-statistic to the distribution obtained by randomly shuffling the group status of the measurements.

Permutation distribution



P-value = $\Pr(|T^\star| \geq |T_{obs}|)$

Small n: Look at all $\binom{n+m}{n}$ possible shuffles

Large n: Look at a sample (w/ repl) of 1000 such shuffles

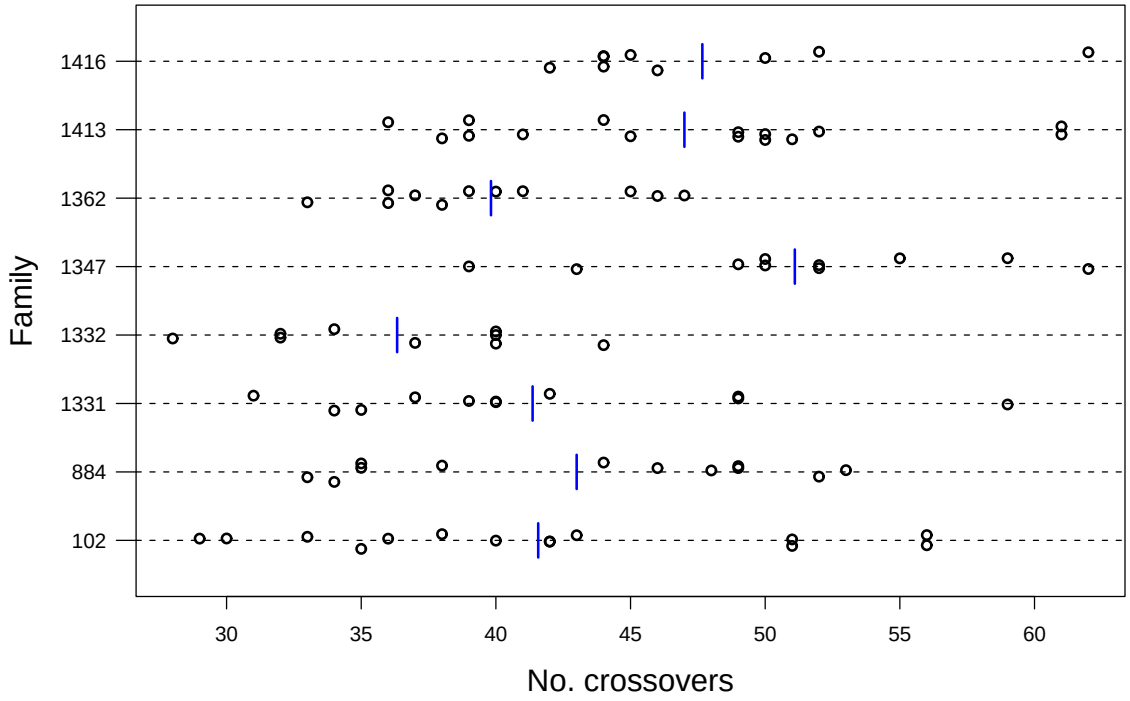
Example

For each of 8 mothers, we observe (estimates of) the number of crossovers, genome-wide, in a set of independent meiotic products.

Question:

Do the mothers vary in the number of crossovers they deliver?

The data



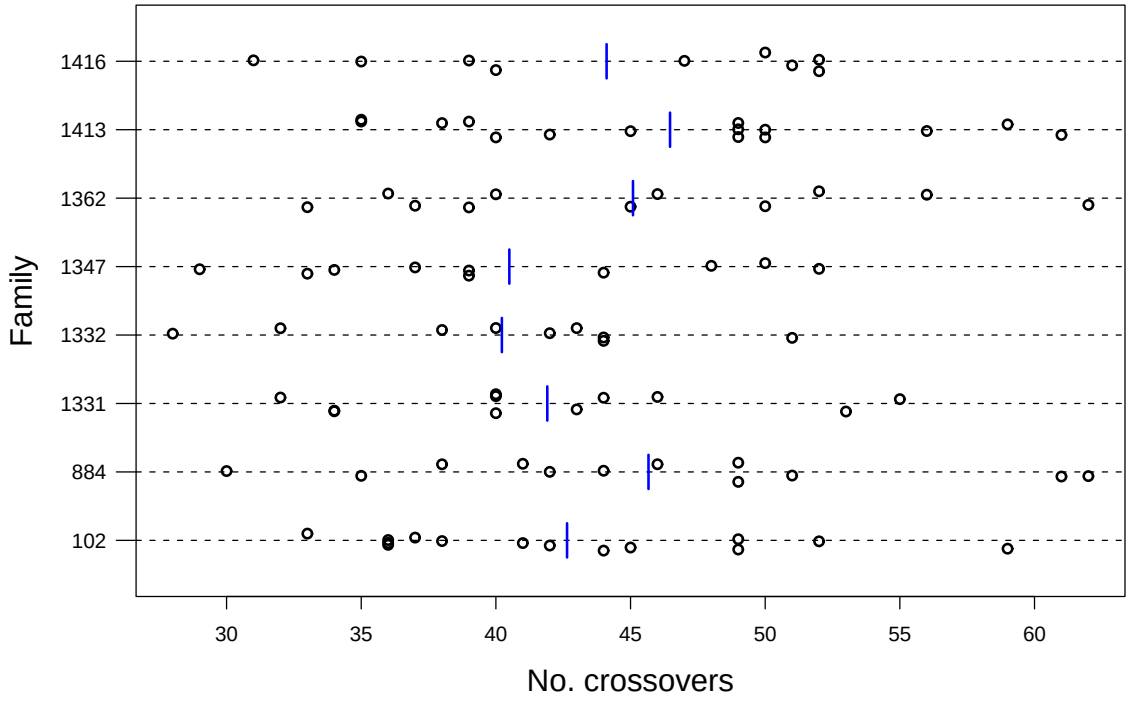
Example (cont.)

How do we think about this?

If there were no relationship between family ID and number of crossovers in a meiotic product:

- What sort of data would we expect?
- What would be the chance of obtaining data as extreme as what was observed?

Permuted data



The F statistic

k groups
n_i values in group i
y_{ij} = jth value in ith group

Assume the data are independent, and that $y_{ij} \sim \text{Normal}(\mu_i, \sigma^2)$.

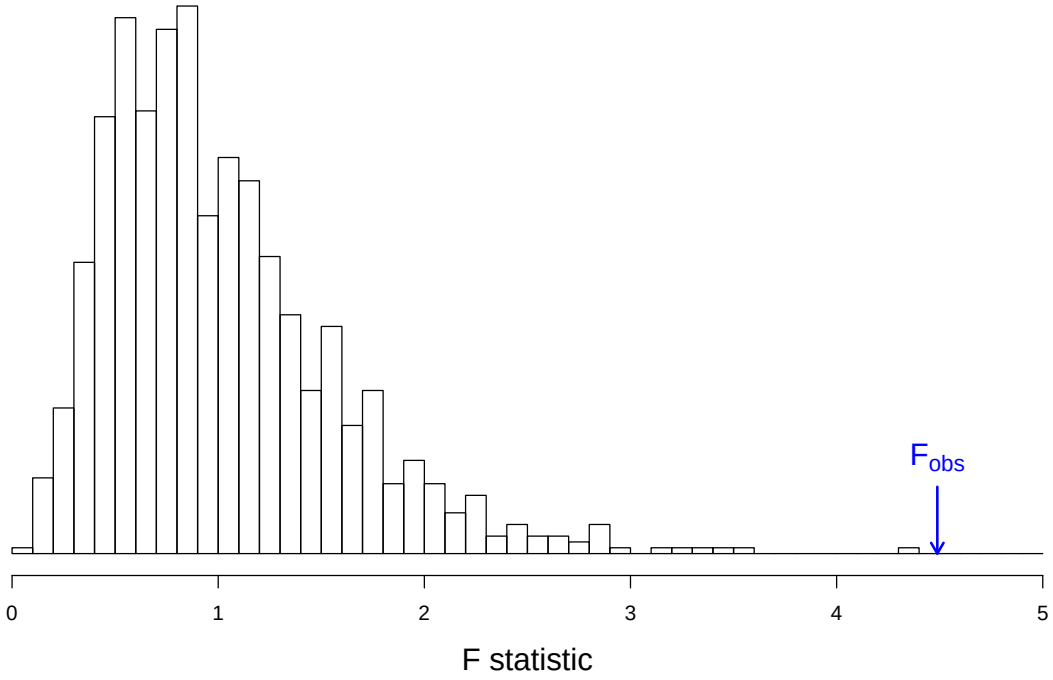
Question: Are the μ_i the same?

Let $\bar{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} / \sum_{i=1}^{k} n_i$

Let $\text{SS}_B = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ and $\text{SS}_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$

$$F = \frac{\text{SS}_B/(k-1)}{\text{SS}_W/\sum(n_i - 1)}$$

Estimated permutation distribution



The above example concerns
the analysis of variance (ANOVA)

Another example

Consider n intercross mice.

Let y_i = phenotype of mouse i

Let $x_i = \begin{cases} 2 & \text{BB} \\ 1 & \text{if mouse } i \text{ is AB at a marker} \\ 0 & \text{AA} \end{cases}$

Imagine $y_i \sim \text{normal}(\mu_{x_i}, \sigma^2)$

Question: $\mu_0 = \mu_1 = \mu_2$?

ANOVA as linear regression

Let $a_i = \begin{cases} +1 & 2 \\ 0 & \text{if } x_i = 1 \\ -1 & 0 \end{cases}$ and $d_i = \begin{cases} 0 & 2 \\ +1 & \text{if } x_i = 1 \\ 0 & 0 \end{cases}$

Then

$$y_i = \mu + \alpha\, a_i + \delta\, d_i + \epsilon_i \qquad \epsilon_i \sim \text{normal}(0, \sigma^2)$$

Estimate μ, α, δ by least squares
(equivalent to maximum likelihood here)

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\mu} - \hat{\alpha}\, a_i - \hat{\delta}\, d_i)^2$$

Note: $\mu_{BB} = \mu + \alpha$ $\mu_{AB} = \mu + \delta$ $\mu_{AA} = \mu - \alpha$

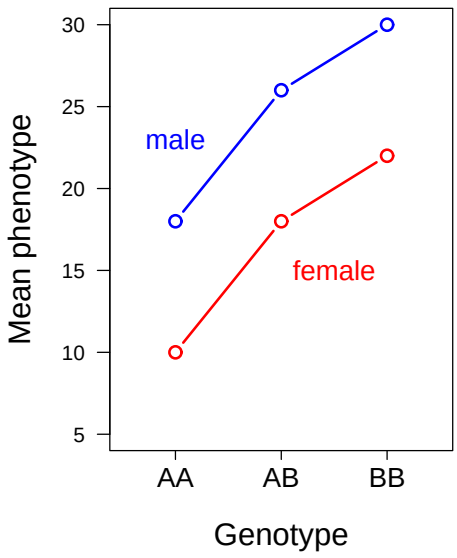
A sex effect

Let s_i = 1/0 if mouse i is male/female.

Consider the model: $y_i = \mu + \alpha\ a_i + \delta\ d_i + \beta\ s_i + \epsilon_i$

	mean phenotype	
marker genotype	females	males
BB	$\mu + \alpha$	$\mu + \beta + \alpha$
AB	$\mu + \delta$	$\mu + \beta + \delta$
AA	$\mu - \alpha$	$\mu + \beta - \alpha$

A sex effect



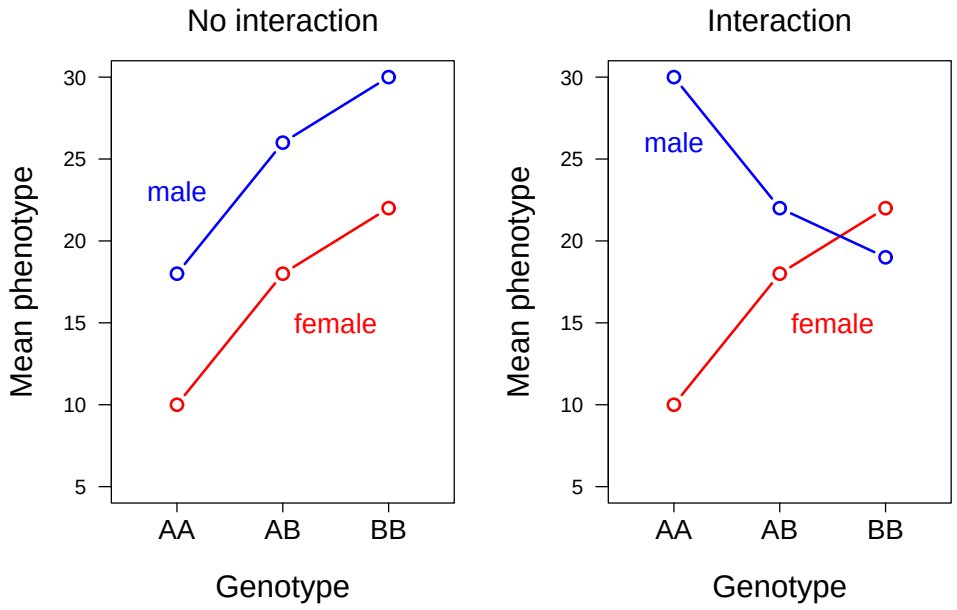
Sex by marker interaction

Consider the model:

$$y_i = \mu + \alpha\ a_i + \delta\ d_i + \beta\ s_i + \gamma_\alpha\ (a_i\ s_i) + \gamma_\delta\ (d_i\ s_i) + \epsilon_i$$

	mean phenotype	
marker genotype	females	males
BB	$\mu + \alpha$	$\mu + \beta + \alpha + \gamma_\alpha$
AB	$\mu + \delta$	$\mu + \beta + \delta + \gamma_\delta$
AA	$\mu - \alpha$	$\mu + \beta - \alpha - \gamma_\alpha$

Sex by marker interaction



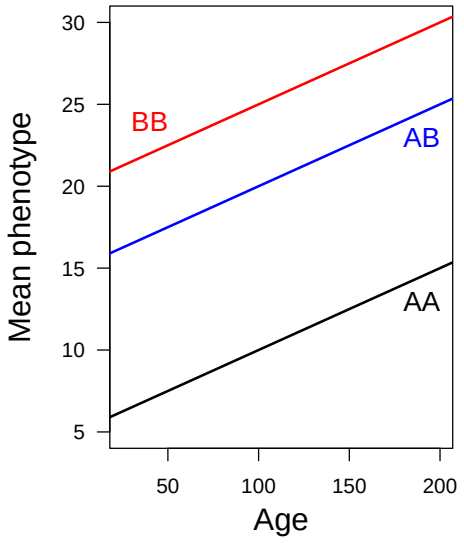
An age effect

Let w_i be the age of mouse i.

Consider the model: $y_i = \mu + \alpha\ a_i + \delta\ d_i + \beta\ w_i + \epsilon_i$

marker genotype	mean phenotype
BB	$\mu + \alpha + \beta\ w_i$
AB	$\mu + \delta + \beta\ w_i$
AA	$\mu - \alpha + \beta\ w_i$

An age effect



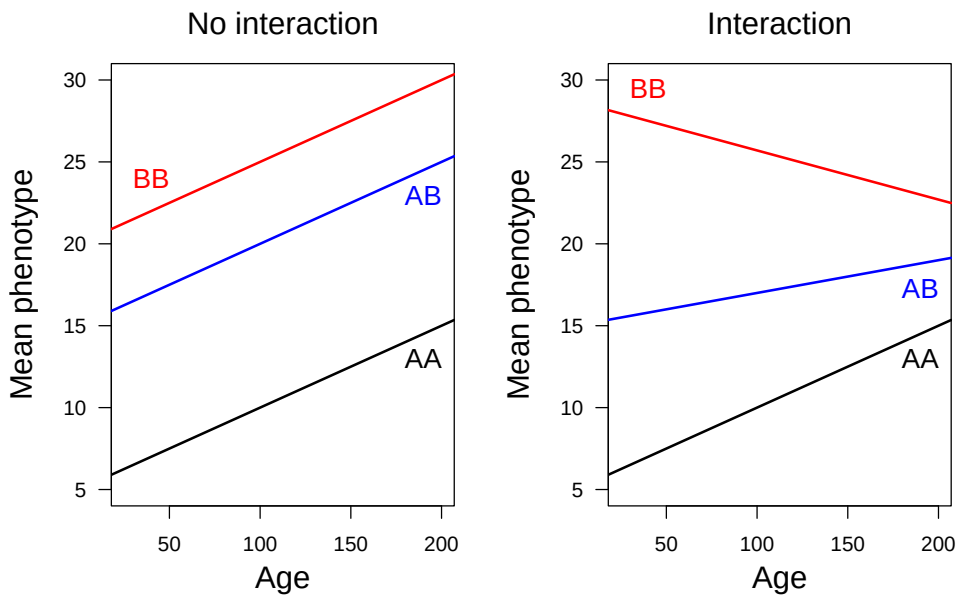
Age by marker interaction

Consider the model:

$$y_i = \mu + \alpha\ a_i + \delta\ d_i + \beta\ w_i + \gamma_\alpha\ (a_i\ w_i) + \gamma_\delta\ (d_i\ w_i) + \epsilon_i$$

marker genotype	mean phenotype
BB	$\mu + \alpha + (\beta + \gamma_\alpha)\ w_i$
AB	$\mu + \delta + (\beta + \gamma_\delta)\ w_i$
AA	$\mu - \alpha + (\beta - \gamma_\alpha)\ w_i$

Age by marker interaction



Frequentist statistics

The target is inference regarding some fixed population parameter, θ.

We imagine repeating the sampling/measurement process.

We focus on the sampling distribution of a statistic/estimator.

Bayesian statistics

Imagine the population parameter, θ, is drawn from some prior distribution, $\Pr(\theta)$.

We calculate the distribution of θ conditional on the observed data (the "posterior" distribution)

$$\Pr(\theta|x) \propto \Pr(x|\theta)\ \Pr(\theta)$$

All inference follows from careful study of $\Pr(\theta|x)$.

Example

Suppose we are interested in the proportion of 8 week old male C57BL/6J mice that survive some infection.

We infect n = 10 mice and note that x = 3 survive.

Frequentist:

$x \sim$ binomial$(10, p)$

A 95% confidence interval for p is 7 – 65%.

In advance, we had a 95% chance that our interval would contain p; after the fact, it either does or doesn't.

Bayesian:

$x \sim$ binomial$(10, p)$

Imagine that $p \sim$ uniform$(0, 1)$, a priori.

There's a 95% chance that p is in the interval 9–59%.

References

- L Gonick, W Smith (1993) *The cartoon guide to statistics*. HarperCollins.
- D Freedman, R Pisani, R Purves (1998) *Statistics*, 3rd edition. Norton.
- JA Rice (1995) *Mathematical statistics and data analysis*, 2nd edition. Duxbury.
- ML Samuels, JA Witmer (1999) *Statistics for the life sciences*, 2nd edition. Prentice Hall.
- FL Ramsey, DW Schafer (2002) *The statistical sleuth: A course in methods of data analysis*, 2nd edition. Duxbury.
- PV Rao (1998) *Statistical research methods in the life sciences*. Duxbury.
- BJF Manly (1997) *Randomization, bootstrap and Monte Carlo methods in biology*, 2nd edition. Chapman & Hall/CRC