Today

- Recap & questions from homework
- Coding likelihood intervals

p-values

- Use constraint; prefer uncertainty intervals
- Key points
  - p-value is not the probability that: “null is true”, “data were generated by the null”, “by chance alone”
  - $p < 0.05$ does not mean “the null hypothesis is false”
  - small p-value does not mean “the effect was large or important”
  - $p > 0.05$ does not mean “there was no effect”, or “the null is true”, or “the effect was small”
  - if many replicated studies have $p > 0.05$ it does not mean “there was no effect”

Likelihood in data science

- This week: pure likelihood inference
  - Learning goal: understand likelihood
- Likelihood is also used in
  - Frequentist: as a sample statistic
  - Bayesian: part of the posterior
  - Information theory: e.g. AIC
    - penalized likelihood

The likelihood function

Counts all the ways the data could have happened for a given model or hypothesis

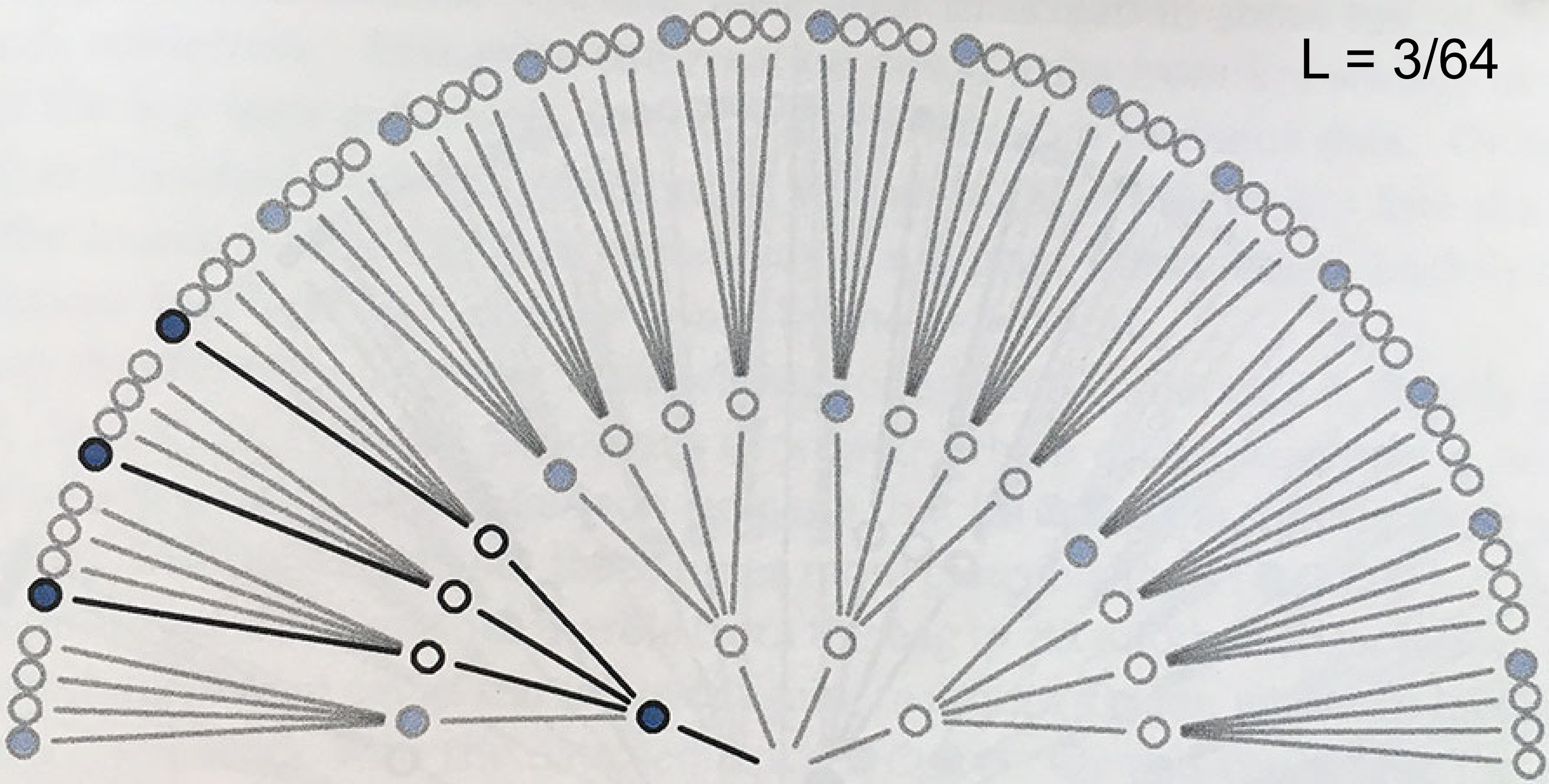


FIGURE 2.3. After eliminating paths inconsistent with the observed sequence, only 3 of the 64 paths remain.

Paths for data  given M_2 

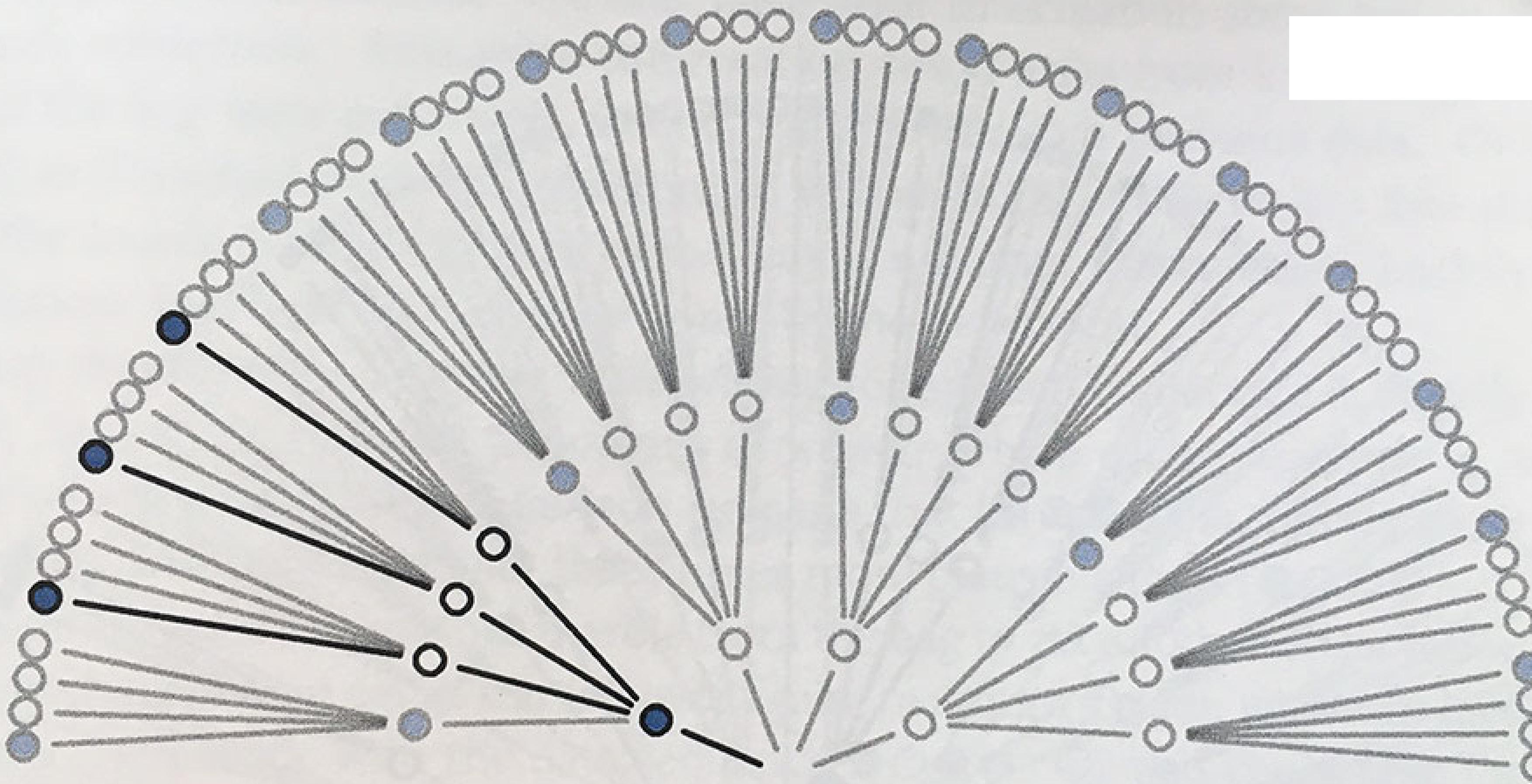
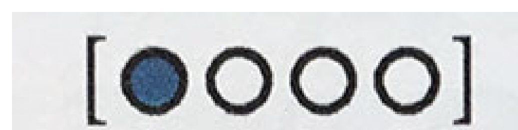


FIGURE 2.3. After eliminating paths inconsistent with the observed sequence, only 3 of the 64 paths remain.

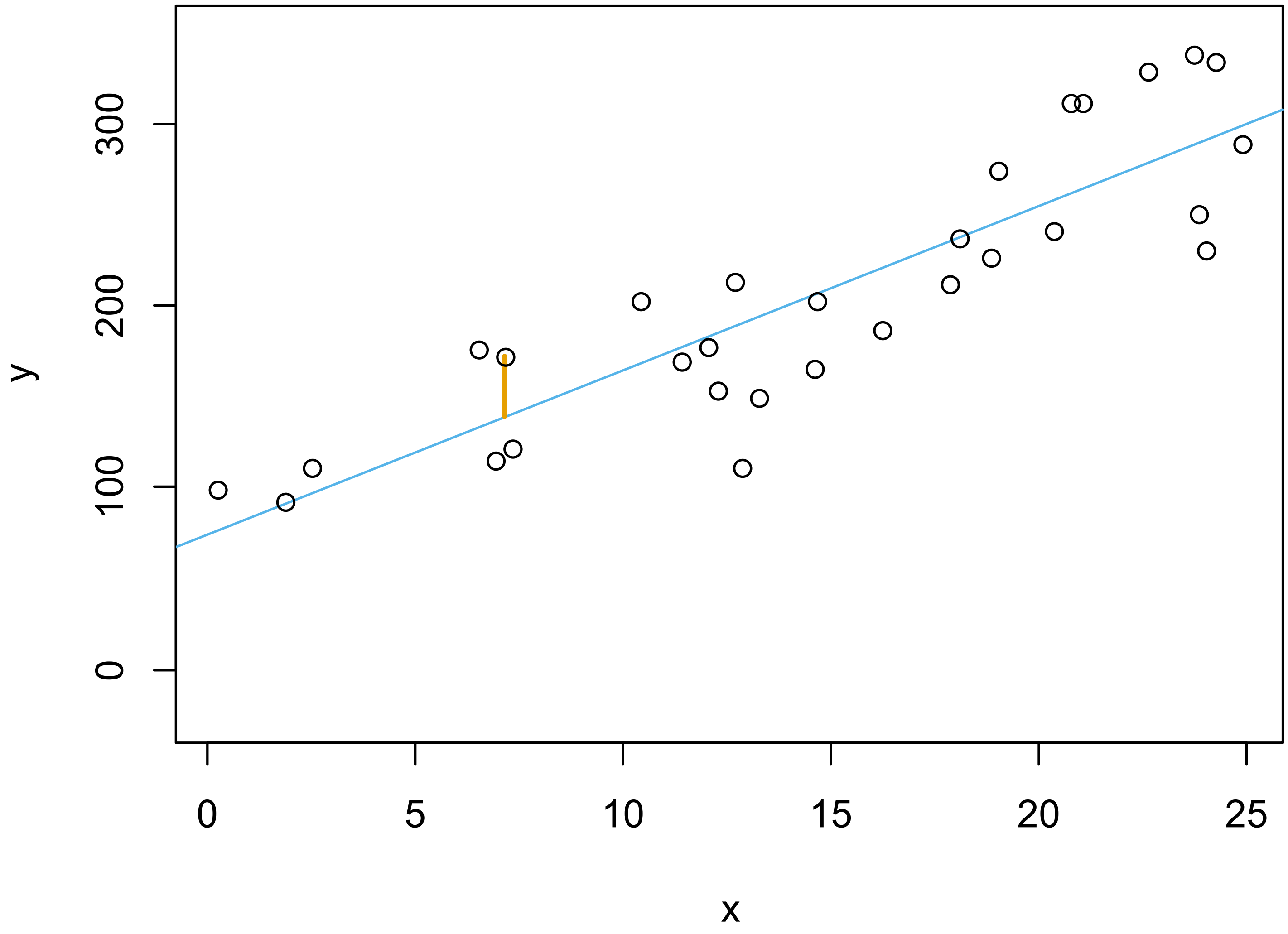
What is the likelihood for
2 blue + 1 white in any order?

given M_2

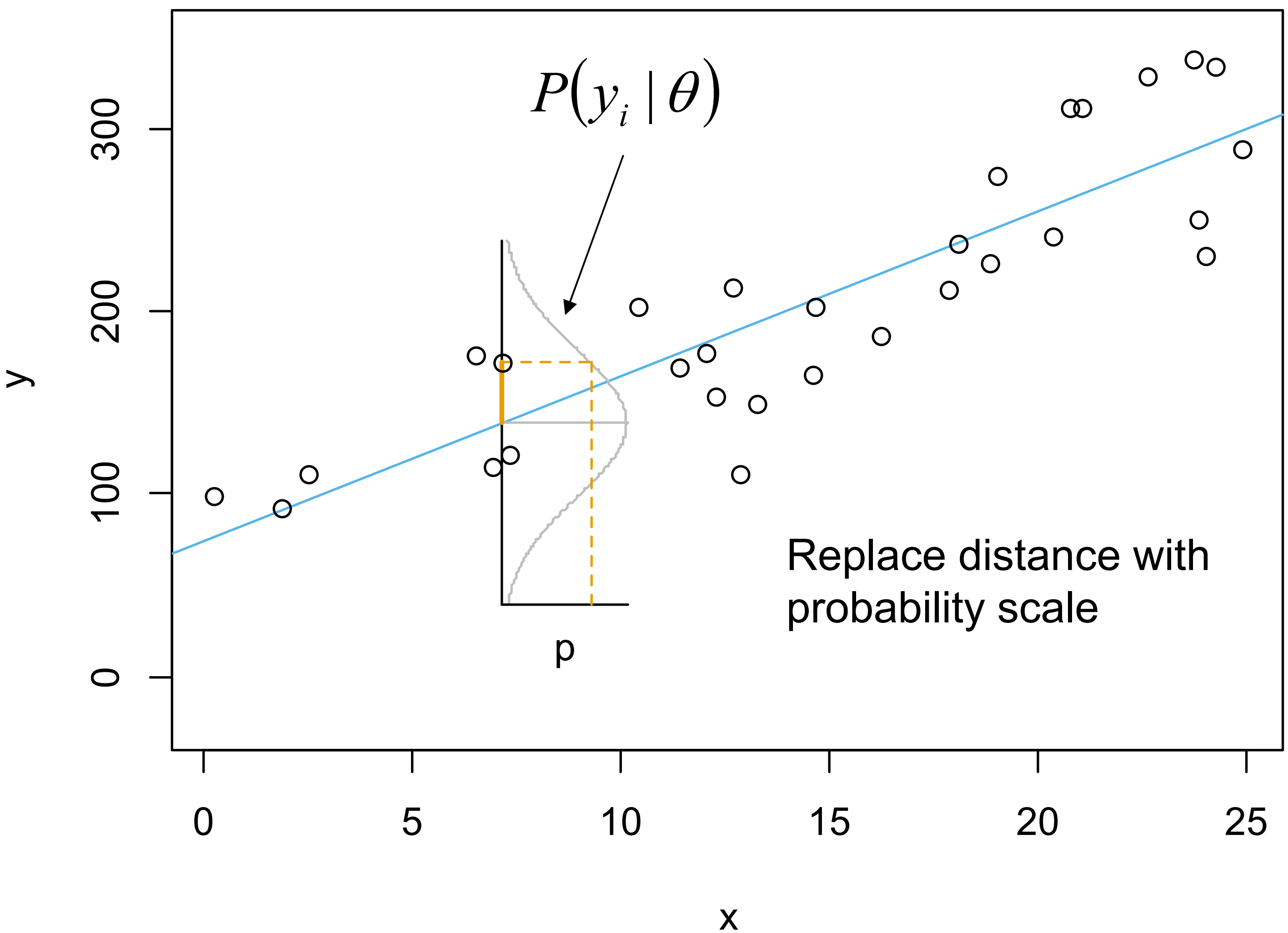


Likelihood inference for the linear model

Likelihood (linear, Normal)



Likelihood (linear, Normal)




```
dnorm(y[i], mean=beta_0 + beta_1 * x[i], sd=sd_pred)
```

Likelihood (linear, Normal)

Writing down the model:

$$y \sim \mathrm{Normal}(\mu, \sigma)$$

$$\mu = \beta_0 + \beta_1 x$$

Likelihood for the model:

$$L(\theta) = P(y \mid \theta) = P(y \mid \beta_0, \beta_1, \sigma)$$

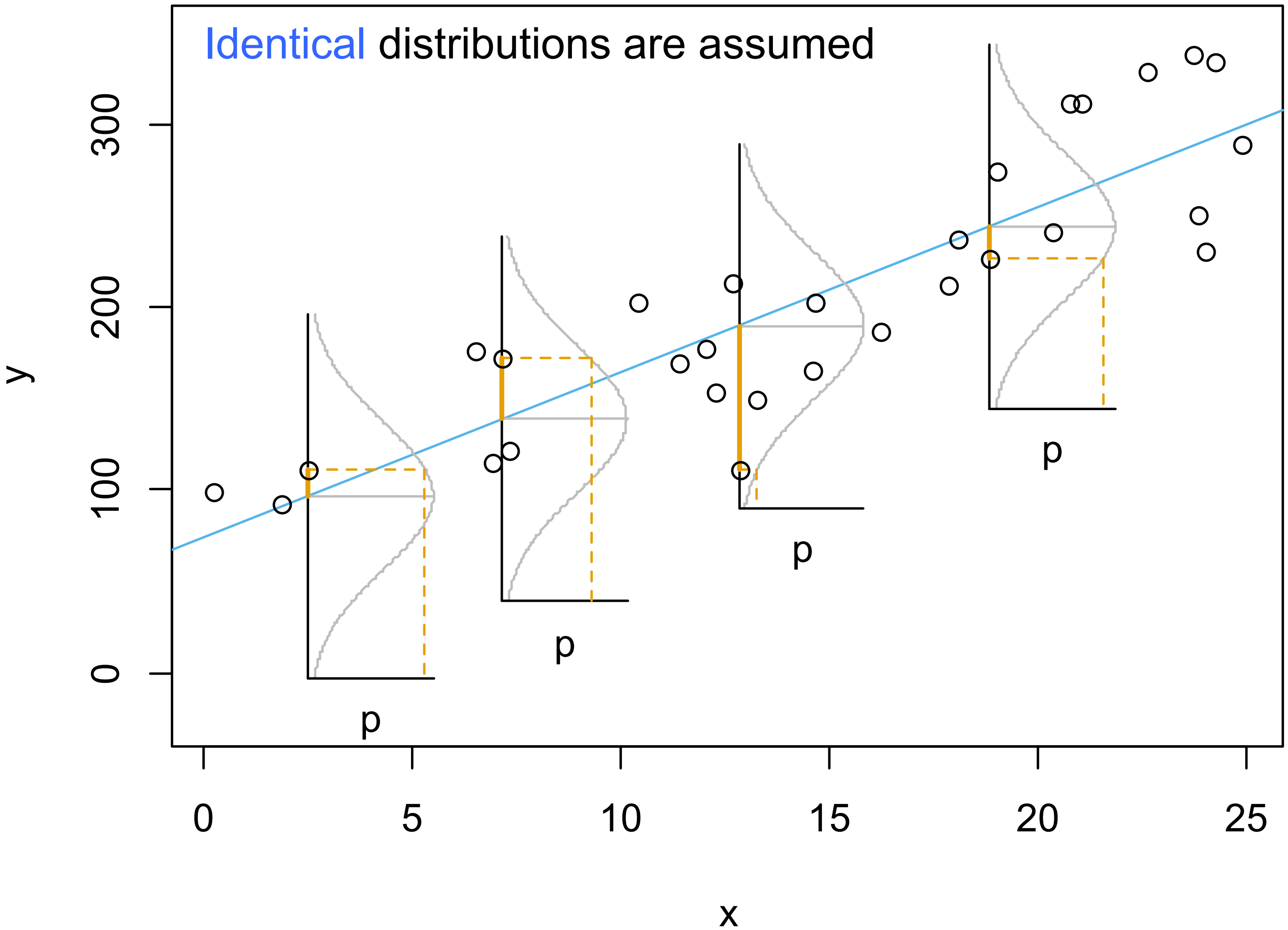
Total likelihood for a data set

One data point: $P(y_1 \mid \theta)$

All data points: $\prod_{i}^{n} P(y_i \mid \theta)$

because probabilities multiply together to give total probability (*n* is the number of datapoints). Independence is assumed.

Likelihood (linear, Normal)



Support function

The log likelihood:

$$\sum_{i}^{n} \ln P\left(y_i \mid \theta\right)$$

Instead of multiplying small probabilities, it is more accurate and convenient to sum their logs.

```
sum(dnorm(y, mean=beta_0 + beta_1 * x, sd=sd_pred, log=TRUE))
```

Training algorithm: Maximum likelihood

The values of the parameters that maximize the likelihood. In other words, the model that maximizes the probability of the data.

An optimization problem.

In practice: minimize the negative log likelihood. The model with the most support, has the smallest negative log likelihood.

Maximum likelihood

- Linear, Normal model, 3 parameters
  - intercept
  - slope
  - standard deviation of Normal
- We find maximum likelihood estimates (MLE) for all 3

Inference algorithm

$$\frac{P(y|\theta_2)}{P(y|\theta_1)}$$ Likelihood ratio

Bayes rule to the rescue:

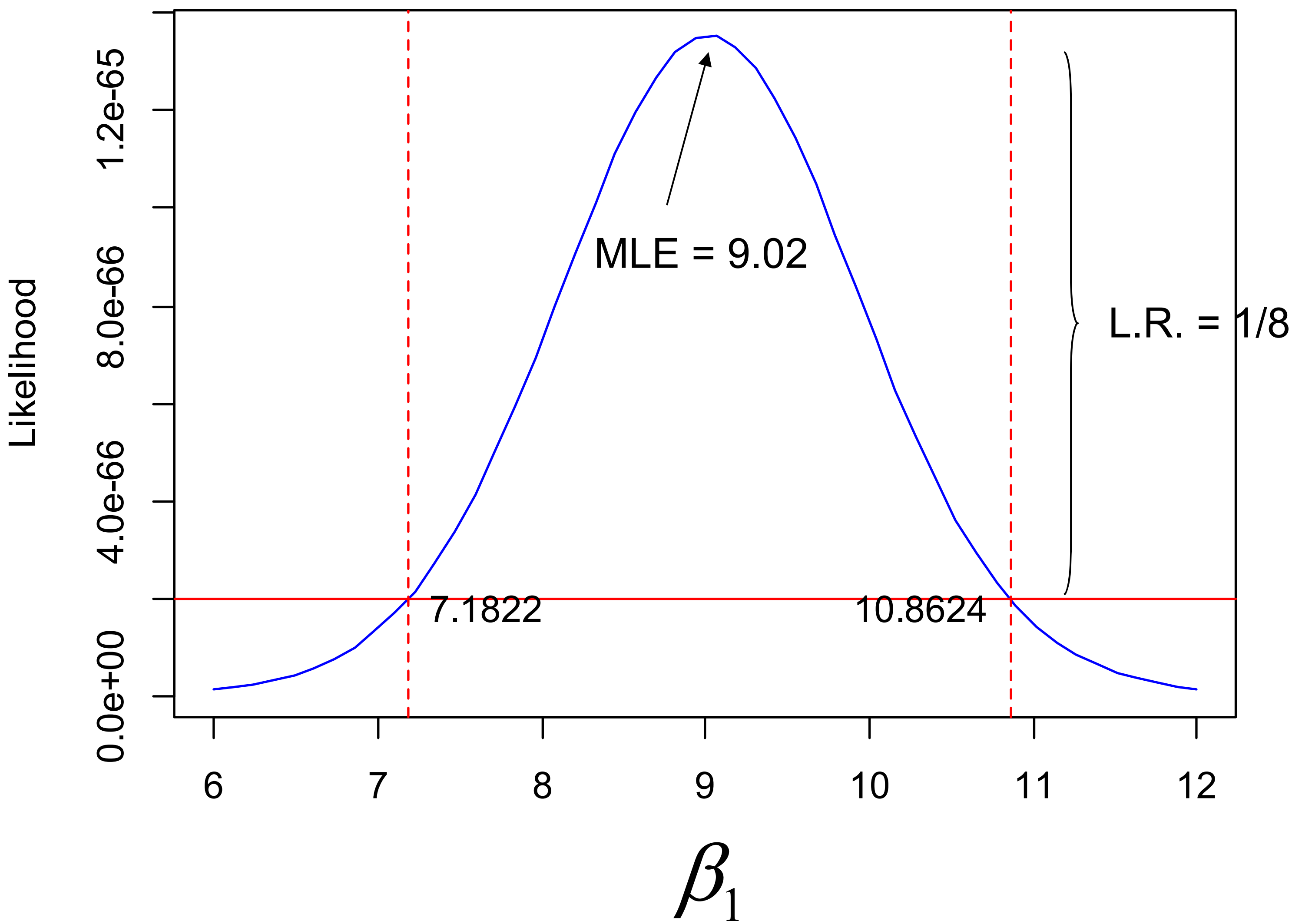
$$\frac{P(\theta_2|y)}{P(\theta_1|y)} = \frac{kP(y|\theta_2)}{kP(y|\theta_1)} = \frac{P(y|\theta_2)}{P(y|\theta_1)} = LR$$

for each pair of models in a set
    calculate likelihood ratio
judge the relative evidence for the models

$$\frac{P(y|\beta_{1i})}{P(y|\beta_{1MLE})}$$

Compare β_1 values for model i against MLE model

Likelihood profile & interval



Calibrating likelihood ratio

- Measure strength of evidence
- How strong do you think it is?
- Two bags with many marbles
  - Bag 1: all white
  - Bag 2: half white, half blue
- 3 whites LR = $1 / 0.5^3 = 2^3 = 8$
- 5 whites LR = $1 / 0.5^5 = 2^5 = 32$
- 10 whites LR = $1 / 0.5^{10} = 2^{10} = 1024$

Compared to SSQ

- Likelihood with a Normal distribution

Likelihood for a dataset

pdf of the Normal distribution

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\cdot\frac{(y_i-\mu_i)^2}{\sigma^2}}$$

y_i are the data points
μ_i is the mean relationship
σ^2 is the variance

Negative log likelihood

$$-\ln\big(L(\theta)\big) = \underbrace{n\left[\ln(\sigma) + \frac{1}{2}\ln(2\pi)\right]} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

This is the SSQ!

Constant w.r.t μ

So, minimizing the nll is the same as minimizing the SSQ

Coding likelihood intervals

- Do it for your data
- Code at end of 06_3_likelihood_inference.Rmd