# ECO 602
# Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

# Today's Agenda

1. First quiz
2. Recap on models
3. Intuition on the descriptive/inferential, sample/population concepts
4. Variables and data types
5. Assignment 1

# Quiz Instructions

Choose 2 questions in the first part.

There aren't right or wrong answers for these questions, but I expect thoughtful replies.

I want to know about your big-picture (conceptual) model thinking process.

If you struggled with the exercise, explain.

**Answer TWO of the following questions – 20 pts:**

1. Describe the question your group decided to ask and your group’s decision process.
2. Which elements of your system were most critical for a model, which could be ignored or simplified?
3. Describe your null and alternative hypotheses.
4. What were some critical unknowns in your system summary and how would address these?

# Why am I harping on model thinking?

1. Model model (not a typo) of data and analyses is sometimes not emphasized enough.
2. When you get into details of a statistical model, it is easy to forget our overall questions.
3. Referring often to your conceptual model of a system helps you remember why you should care about your statistical model.

# Descriptive/inferential, sample/population concepts

Parallel concepts:

1. We can calculate **descriptive** stats on a sample
2. We must **infer** parameters for a larger population

# Statistical populations are context dependent

Depends on system, research questions

Bird example from McGarigal:

Testimony 1, 2

- population = single mountain, unit = single nesting site

Testimony 3, 4:

- population = species range, unit = nesting site

# Samples and sampling units

Sample = **group of observations** taken from larger population.

Sampling unit = ‘thing’ of interest to research question

1. Sampling units are highly context-dependent

Variable = attribute of sampling unit

# Sampling units are context dependent: hypothetical fish size example

Question: characterize the length distribution of 1 species in one lake:

- Statistical population = all fish in the lake
- Ecological population = entire range of the fish species.
- Sampling unit = individual fish
- Variable = fish length
  - What are the units of measurement and the scale of the variable?

# Sampling units are context dependent: hypothetical fish size example

Question: characterize the length distribution of 1 species in MA:

- Statistical population = all fish in MA (in all lakes)
- Ecological population = entire range of the fish species.
- Sampling unit = individual fish
- Variable = fish length

# Sampling units are context dependent: hypothetical fish size example

Question: characterize the length distribution of 1 species in MA

- Statistical population = all fish in all MA (in all lakes)
- Ecological population = entire range of the fish species.
- Sampling unit = individual fish
- Variable = fish length

# Sampling units are context dependent: hypothetical fish size example

Question: characterize the variability fish length among lakes in MA

- Statistical population = all lakes in MA
- Ecological population = entire range of the fish species.
- Sampling Unit = individual lakes
- Variable = mean fish length in the lake

# Descriptive and Inferential: 2 points of view

Population and sample point of view

- We use a **small** sample to learn about the **larger** population

Uncertainty point of view

- We can calculate sample descriptive exactly*.
  - *except for measurement error
- Inference, i.e. estimation, of population parameters introduces uncertainty.

# Quick detour: some numerical and graphical descriptive stats

# Numerical stats: 5 number summary

1. Min
2. Max
3. Median
4. 1st quartile, i.e. 25th percentile
5. 3rd quartile, i.e. 75th percentile

# Graphical summaries

- Scatterplot: individual data points
  - x and y axes: predictor and response
  - shows relationships between variables
- Histogram: counts
  - x axis = bins
  - y axis = count of points in bin
- Boxplot: 5 number summary
  - extreme data points

# Samples and populations: intuition

Amount of variation in population

- Spread is harder to quantify than center

Sample size

- Larger sample = better population estimates, but…
- Diminishing returns
  - Square root terms in formulas

# Hypothetical example population: 10 million mosquitoes

1. Sampling unit: individual mosquitoes
2. Variable of interest: wingspan
3. Scale: continuous, ratio, millimeters

# Hypothetical example population: 10 million mosquitoes

In the real world, we couldn't measure all individuals, but we can create a simulated population.

We can build a simulated population of 10 million using software

# Hypothetical example population: 10 million mosquitoes

Our mosquito population:

- Mean wingspan = 100mm
- Standard deviation of wingspan = 10mm
- Wingspan is Normally* distributed
  - Symmetrical, bell shaped

These are big mosquitoes!

* 'Normal' is usually capitalized when referring to the Normal distribution

# Hypothetical example population: descriptive stats: center, spread, symmetry

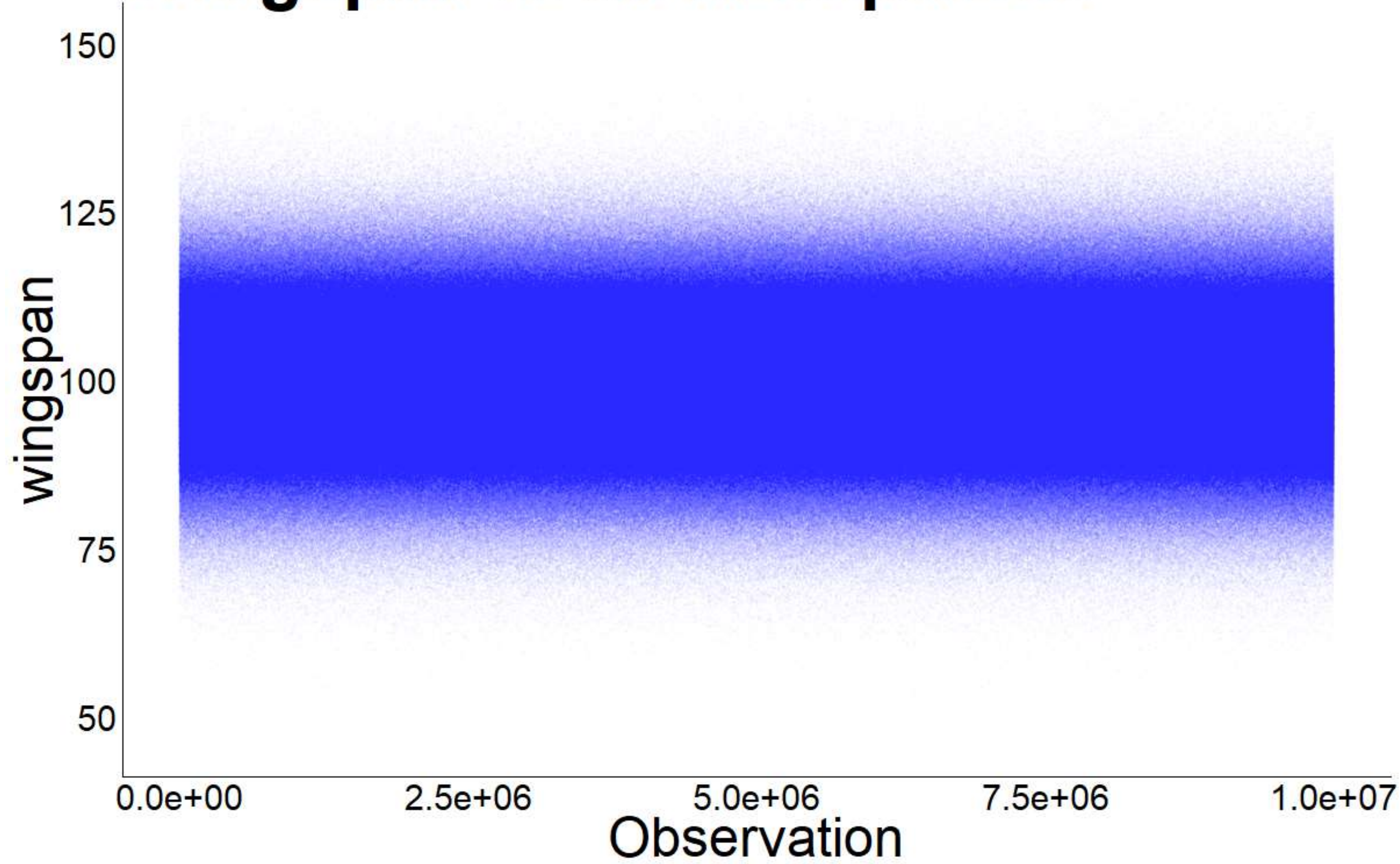
Graphical summaries:

1. Scatterplot
2. Histogram
3. Boxplot

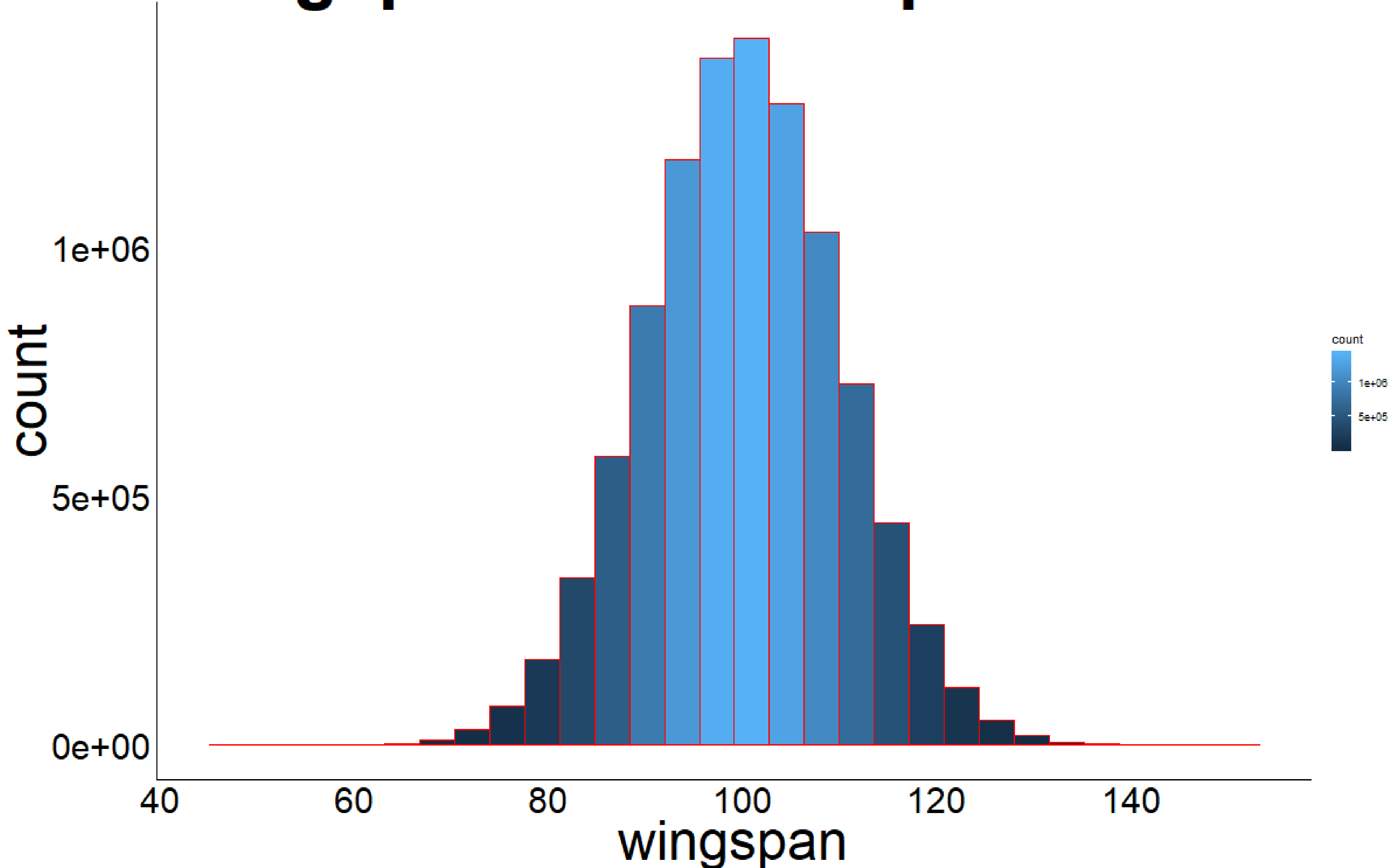
Numerical summaries – descriptive stats:

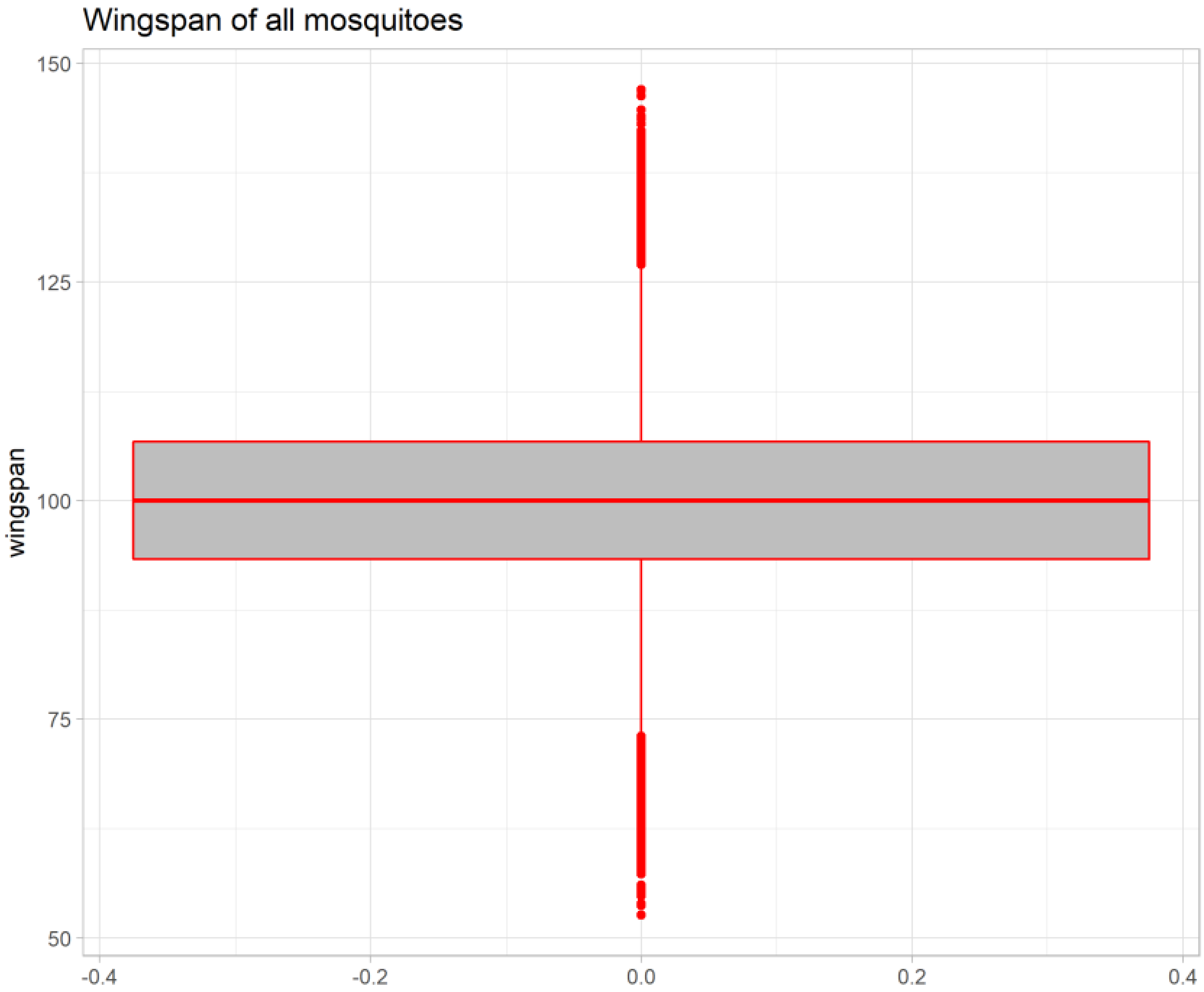
1. Mean
2. Quartiles (percentiles)
3. Standard deviation

# Wingspan of all mosquitoes



# Wingspan of all mosquitoes




Wingspan of all mosquitoes
wingspan
150
125
100
75
50
-0.4
-0.2
0.0
0.2
0.4

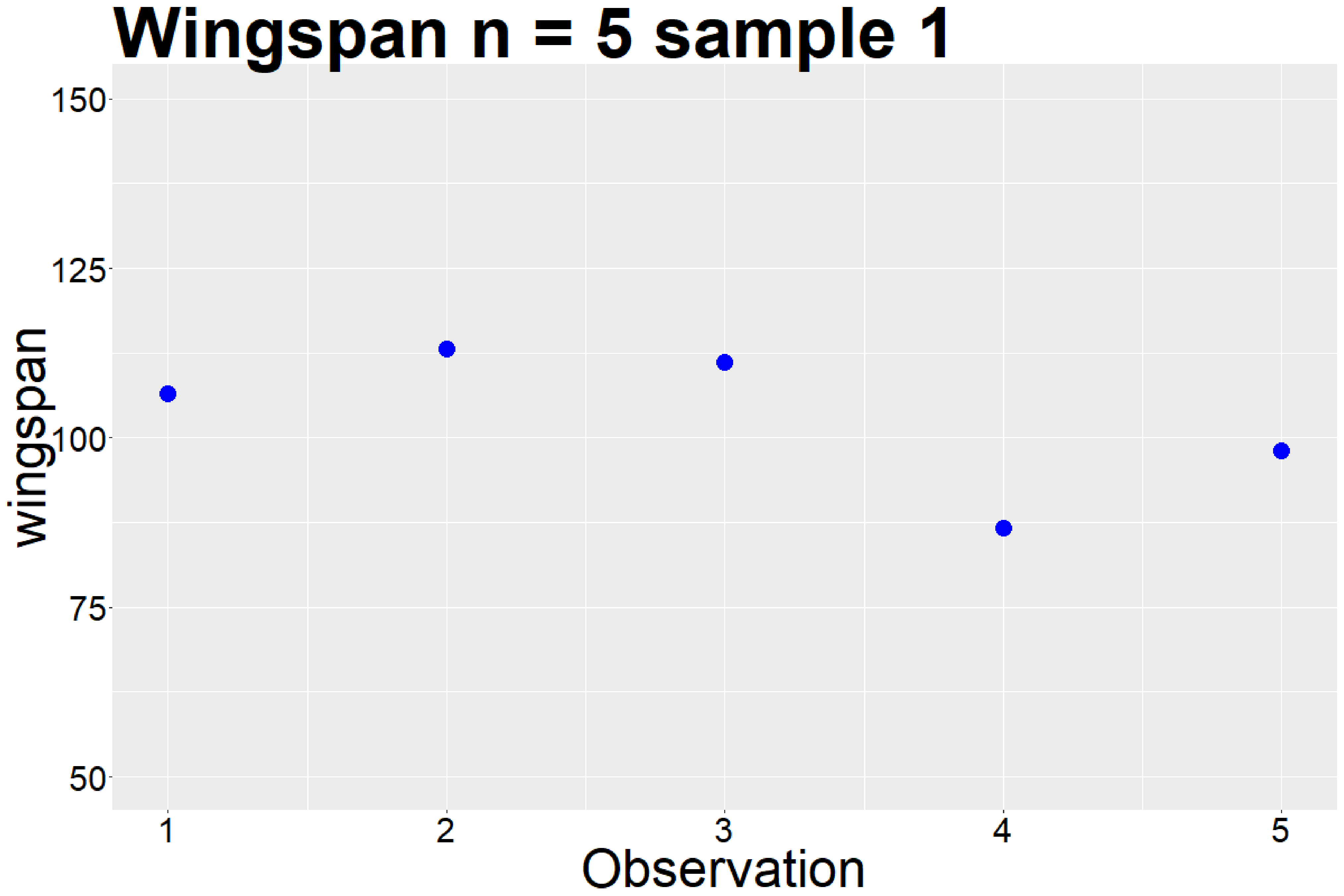
# Hypothetical example population: numerical summaries

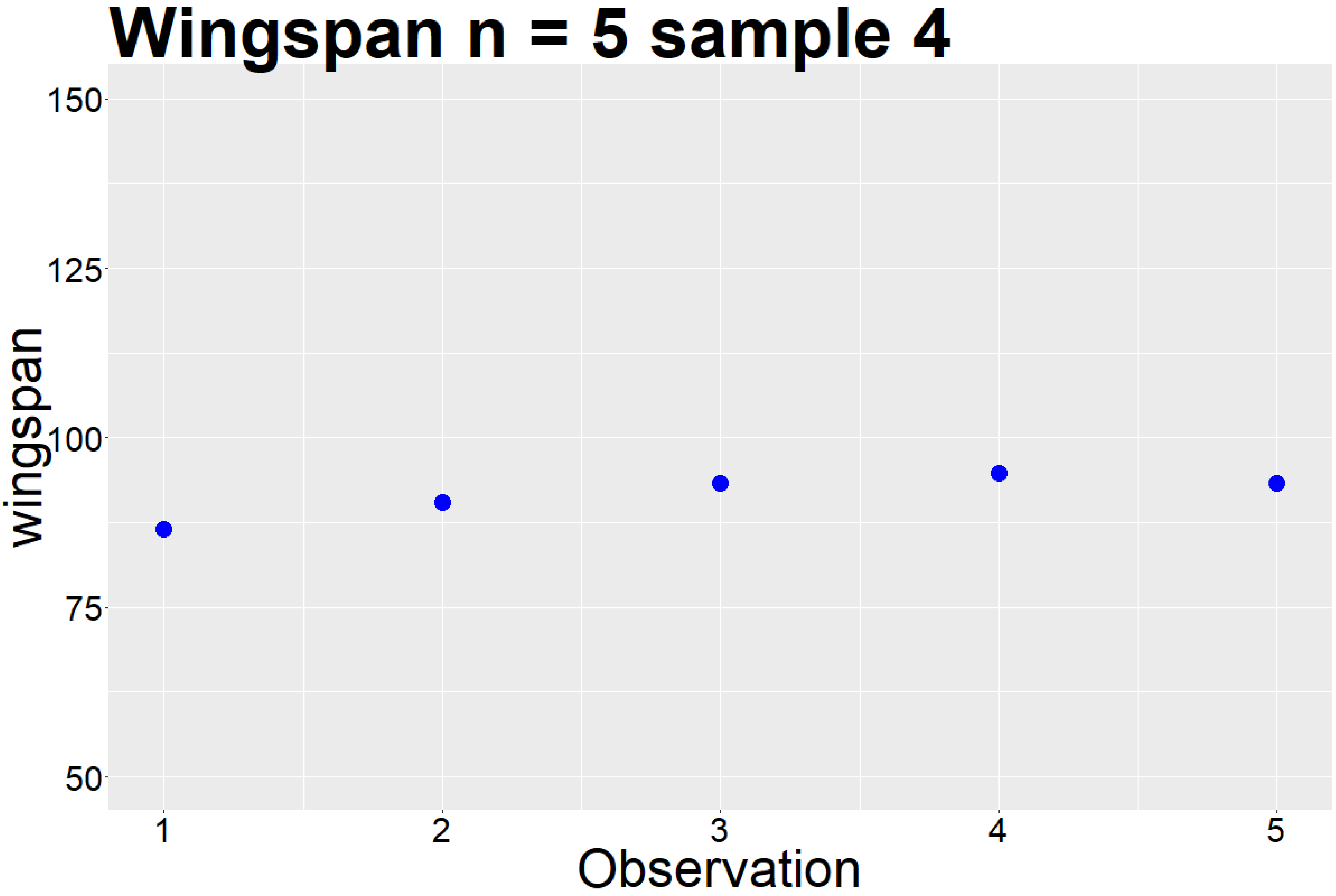
| | Sample | N | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | Whole Population | 10000000 | 100 | 10 | 49.04 | 152.6 |

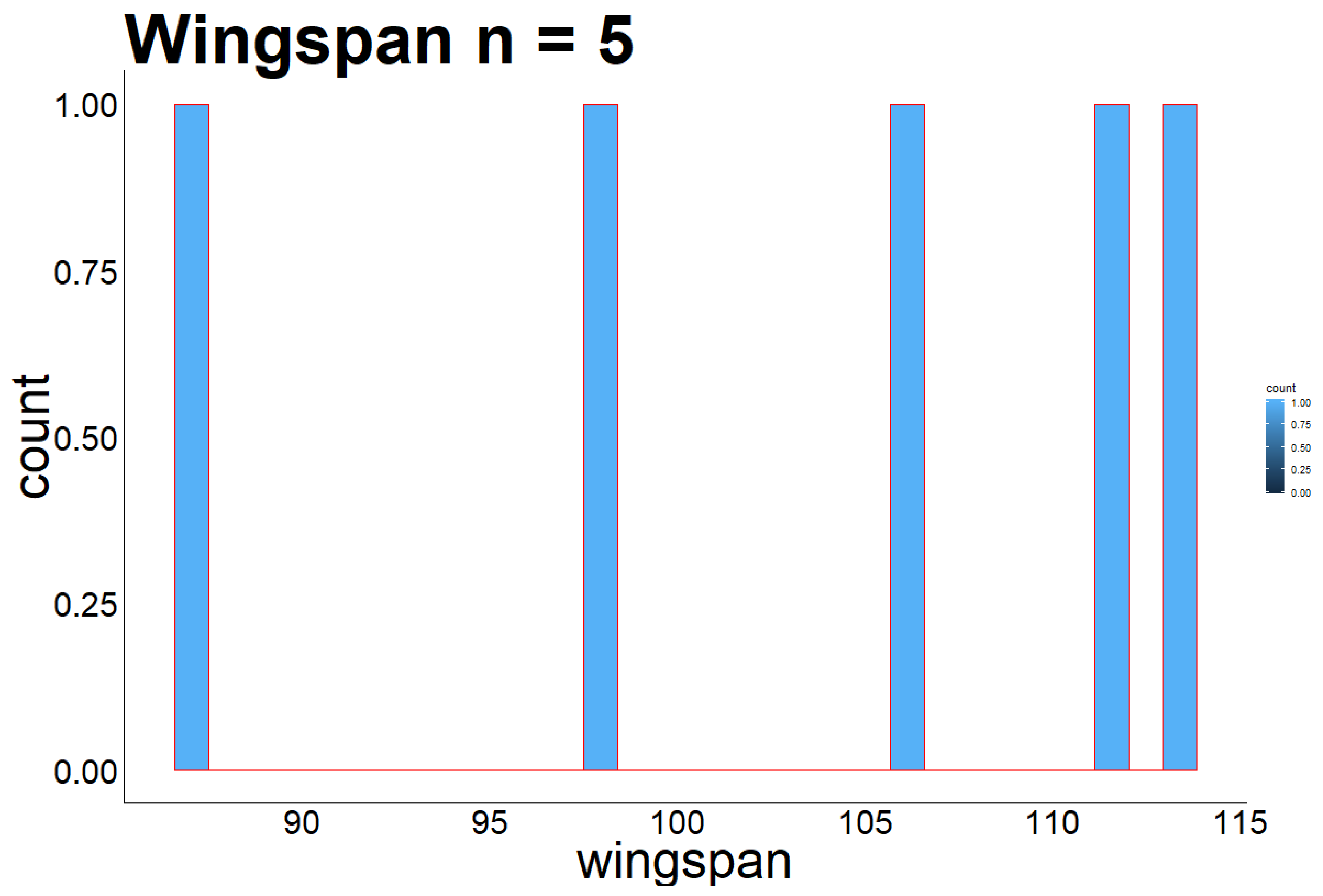
# Sampling: intuition

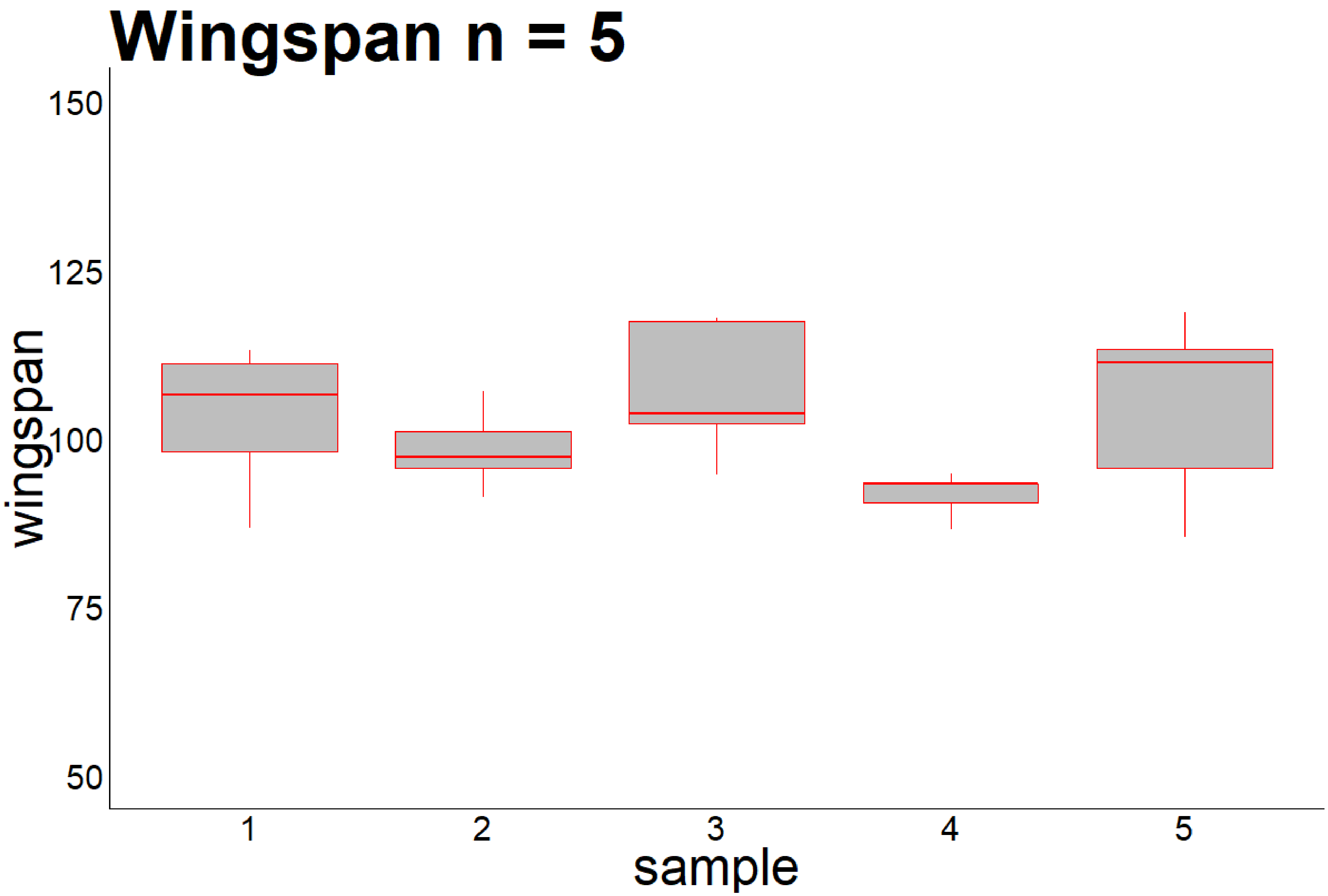
Larger samples = better population estimates

You have a very small budget; you can only afford to sample 5 mosquitoes: $n = 5$


Wingspan n = 5 sample 1
150
125
100
75
50
wingspan
1
2
3
4
5
Observation


Wingspan n = 5 sample 4
150
125
100
75
50
wingspan
1
2
3
4
5
Observation

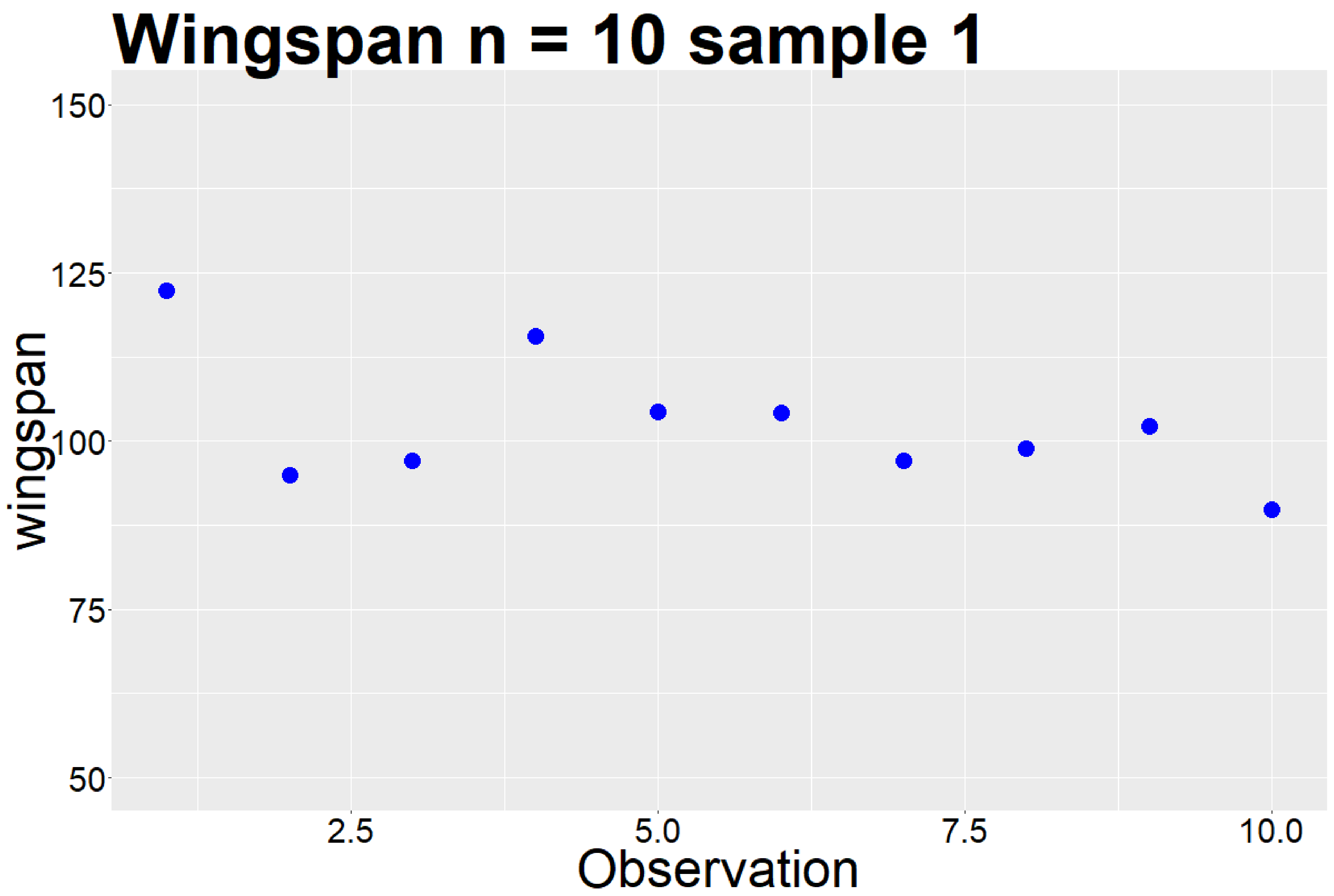

Wingspan n = 5
count
1.00
0.75
0.50
0.25
0.00
90
95
100
105
110
115
wingspan
count
1.00
0.75
0.50
0.25
0.00

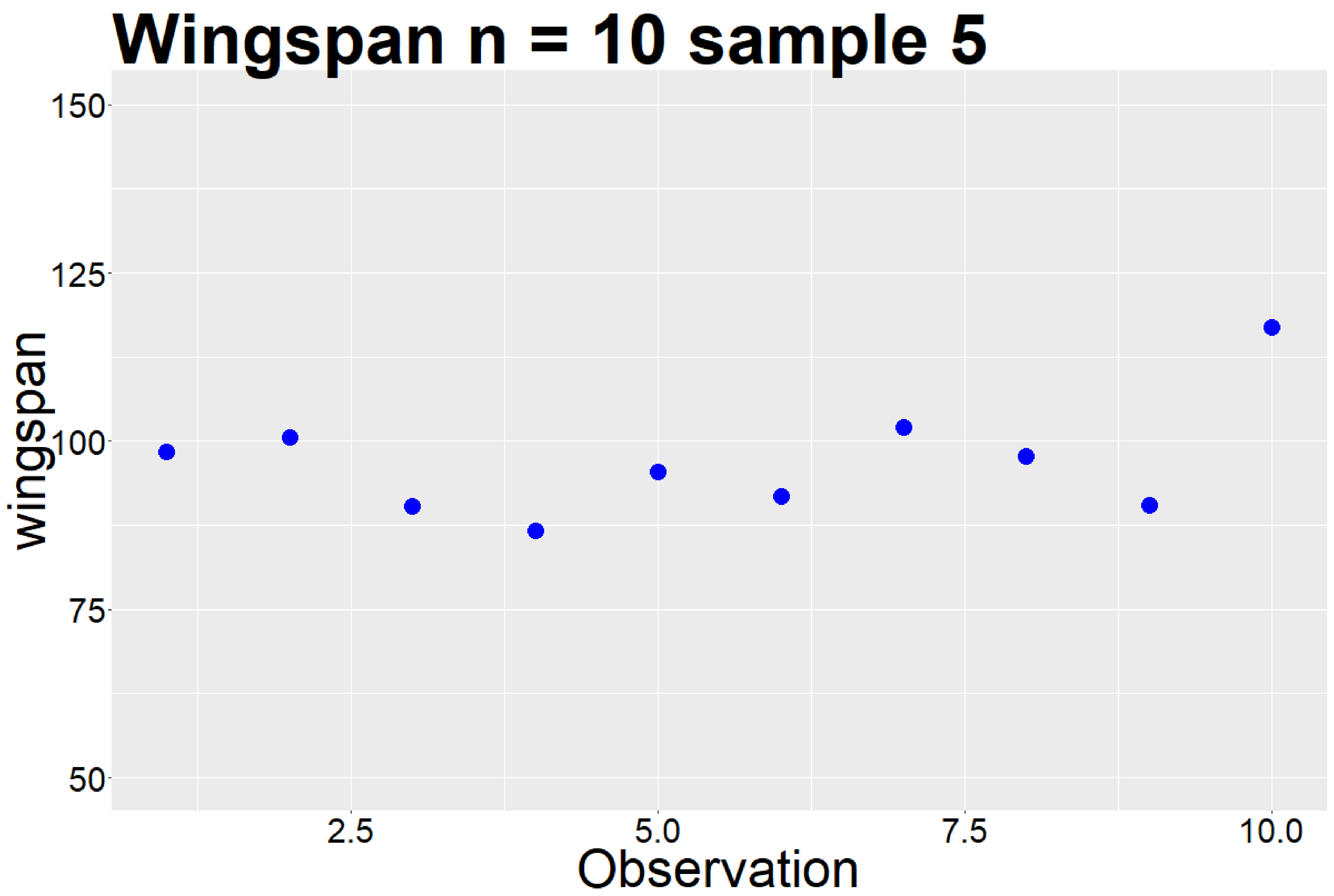

Wingspan n = 5
wingspan
150
125
100
75
50
sample
1
2
3
4
5

| | Sample | n | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | Whole Population | 10000000 | 100 | 10 | 49.04 | 152.6 |
| 2 | Sample 1 | 5 | 103.1 | 10.83 | 86.76 | 113.1 |
| 3 | Sample 2 | 5 | 98.44 | 5.92 | 91.31 | 107 |
| 4 | Sample 3 | 5 | 107.1 | 10.21 | 94.57 | 118 |
| 5 | Sample 4 | 5 | 91.67 | 3.268 | 86.55 | 94.81 |
| 6 | Sample 5 | 5 | 104.8 | 13.87 | 85.37 | 118.7 |

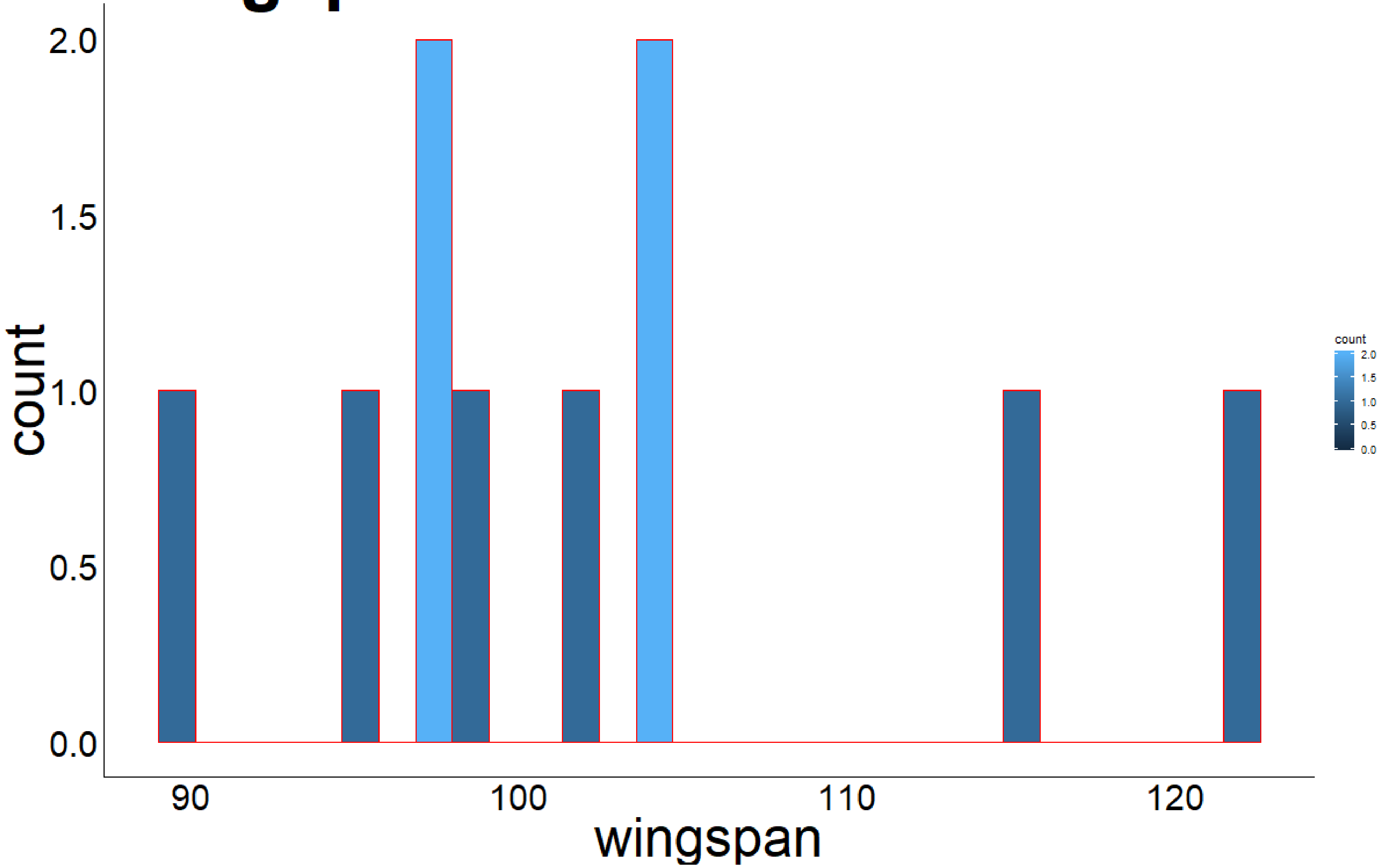
# Sampling: intuition

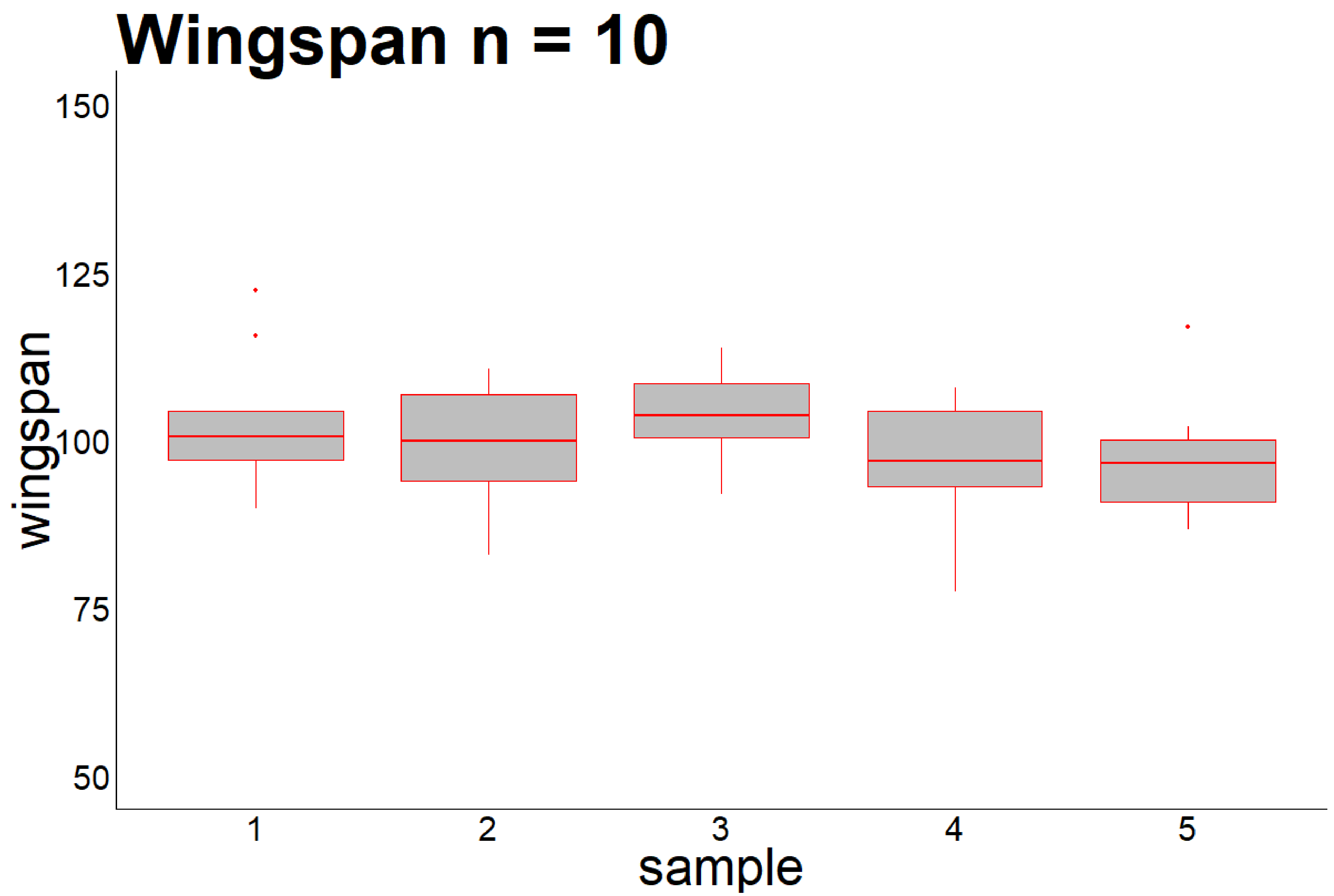
Slightly larger sample: $n = 10$


Wingspan n = 10 sample 1
150
125
100
75
50
wingspan
2.5
5.0
7.5
10.0
Observation


Wingspan n = 10 sample 5
150
125
100
75
50
wingspan
2.5
5.0
7.5
10.0
Observation

# Wingspan n = 10

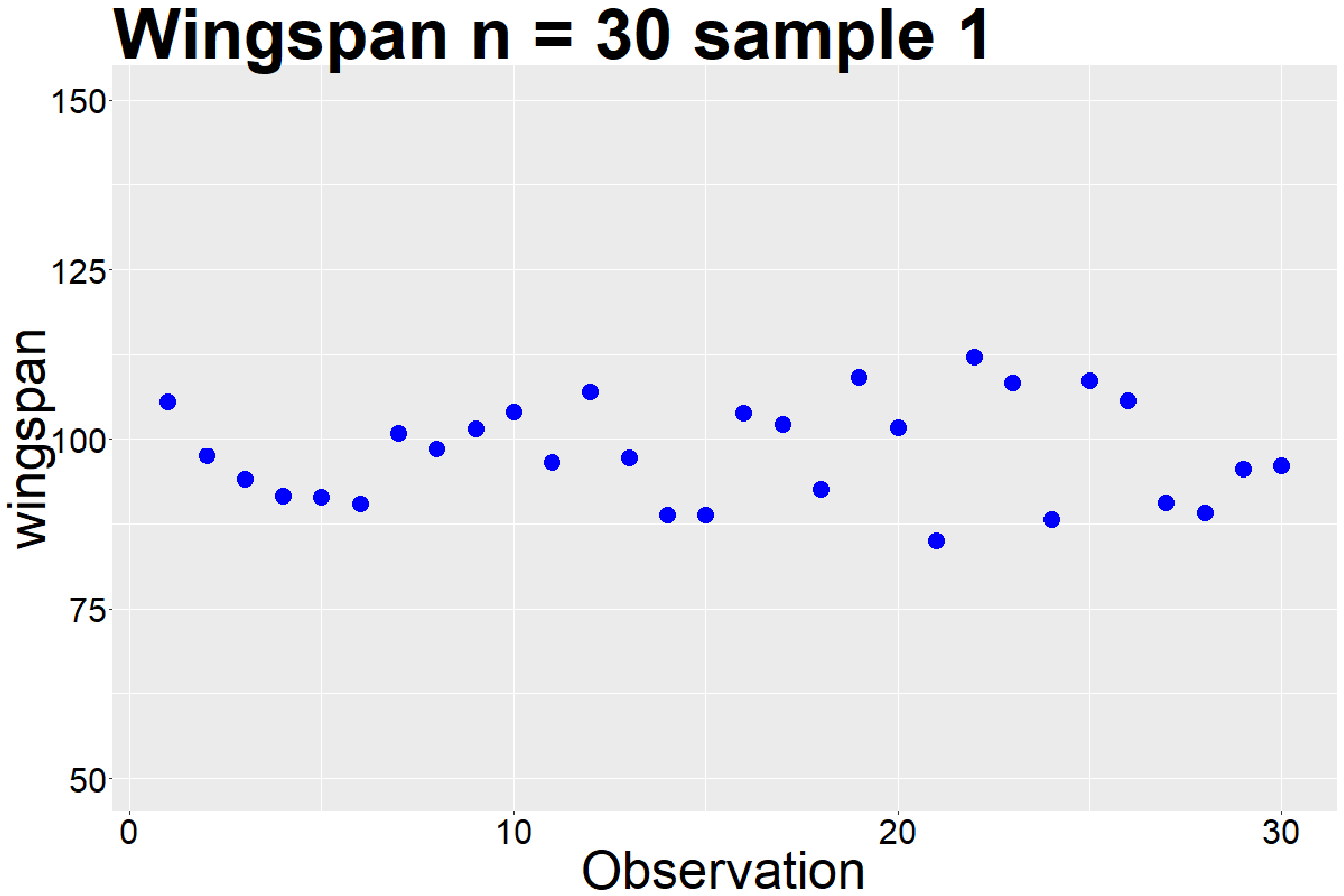



Wingspan n = 10
150
125
100
75
50
wingspan
1
2
3
4
5
sample

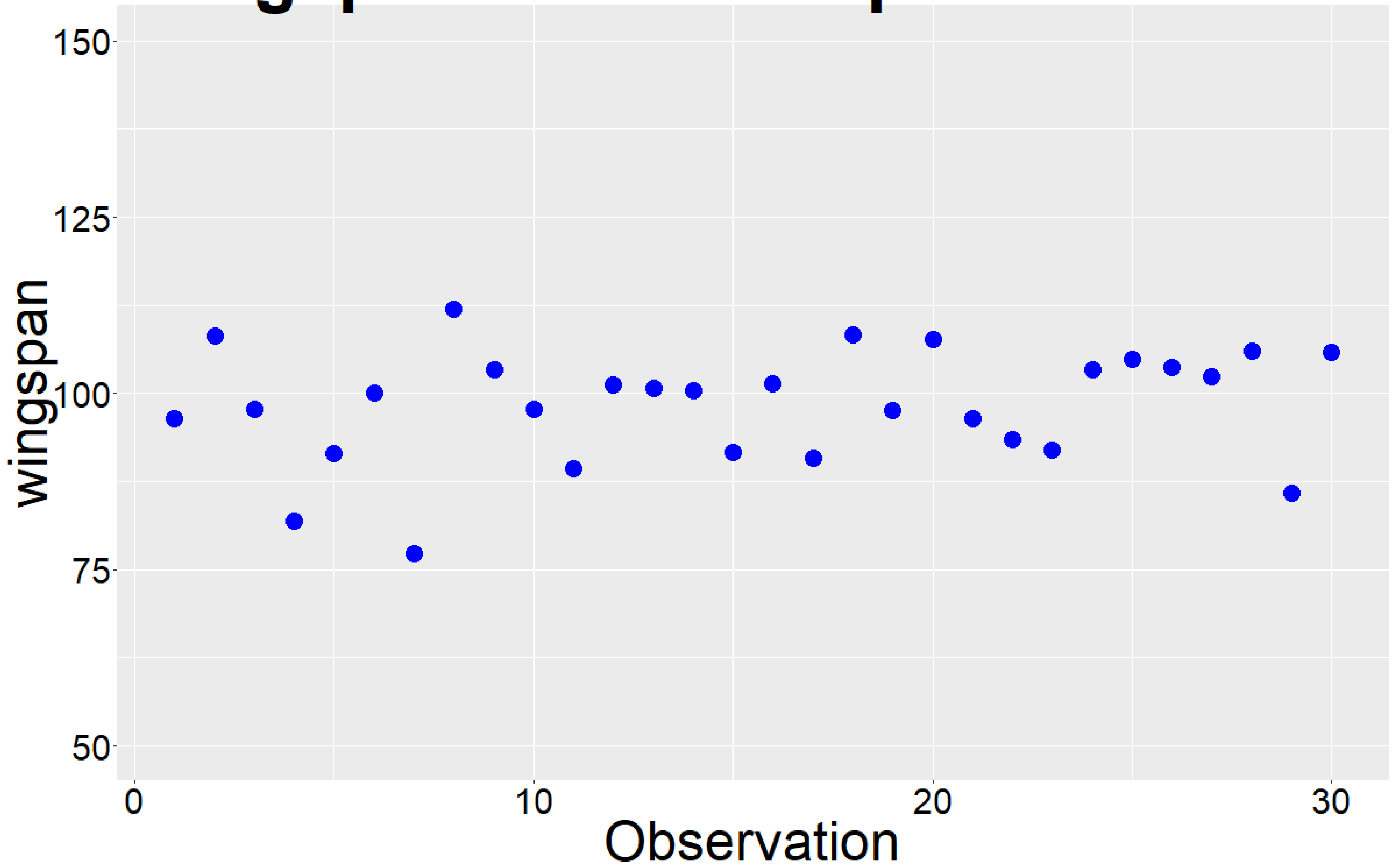
| | Sample | n | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | Whole Population | 10000000 | 100 | 10 | 49.04 | 152.6 |
| 2 | Sample 6 | 10 | 102.7 | 9.757 | 89.86 | 122.3 |
| 3 | Sample 7 | 10 | 99.24 | 9.19 | 82.94 | 110.7 |
| 4 | Sample 8 | 10 | 104 | 6.317 | 91.9 | 113.8 |
| 5 | Sample 9 | 10 | 97.26 | 8.96 | 77.39 | 107.9 |
| 6 | Sample 10 | 10 | 97.06 | 8.554 | 86.71 | 116.9 |

# Sampling: intuition

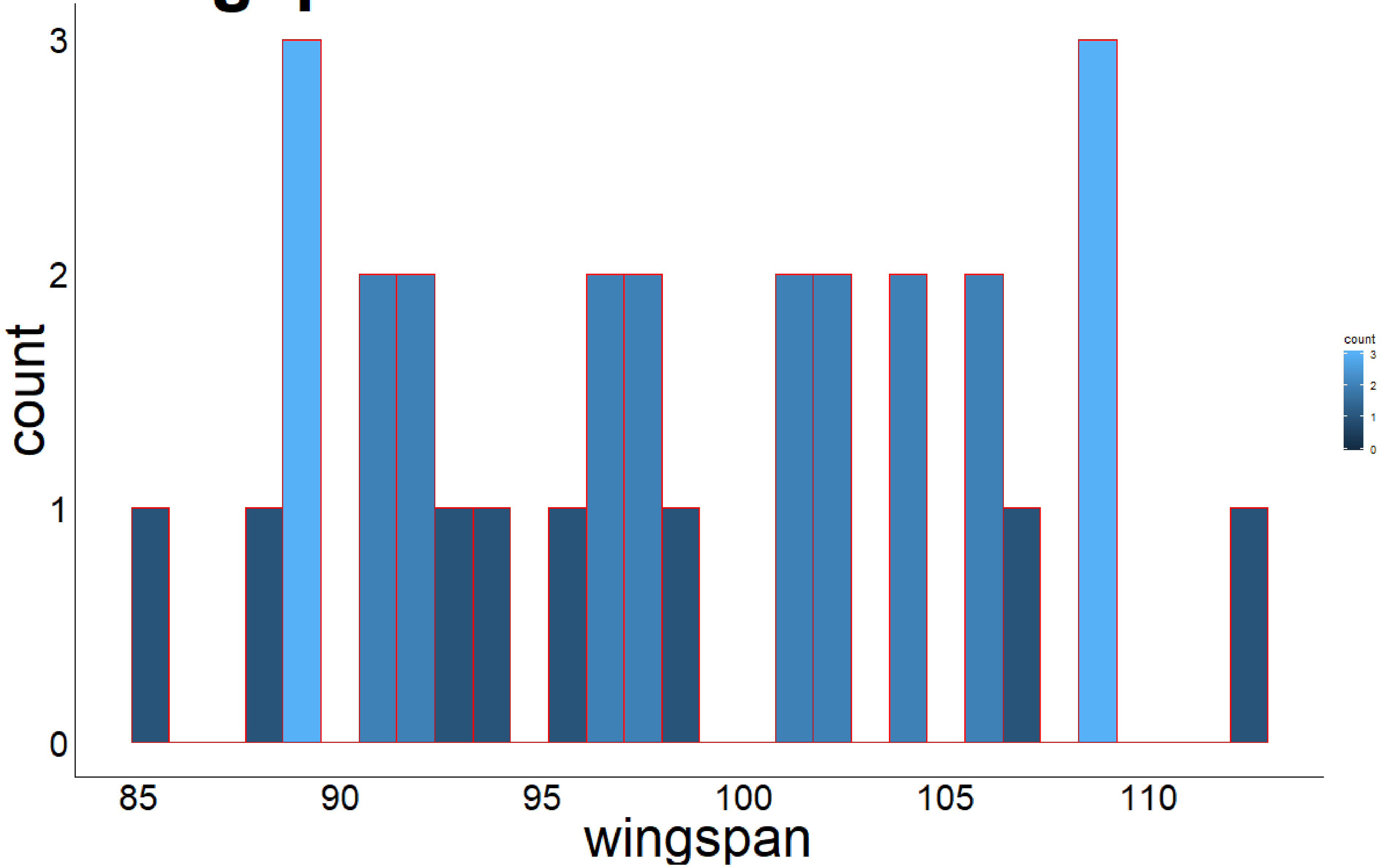
Sample size: $n = 30$

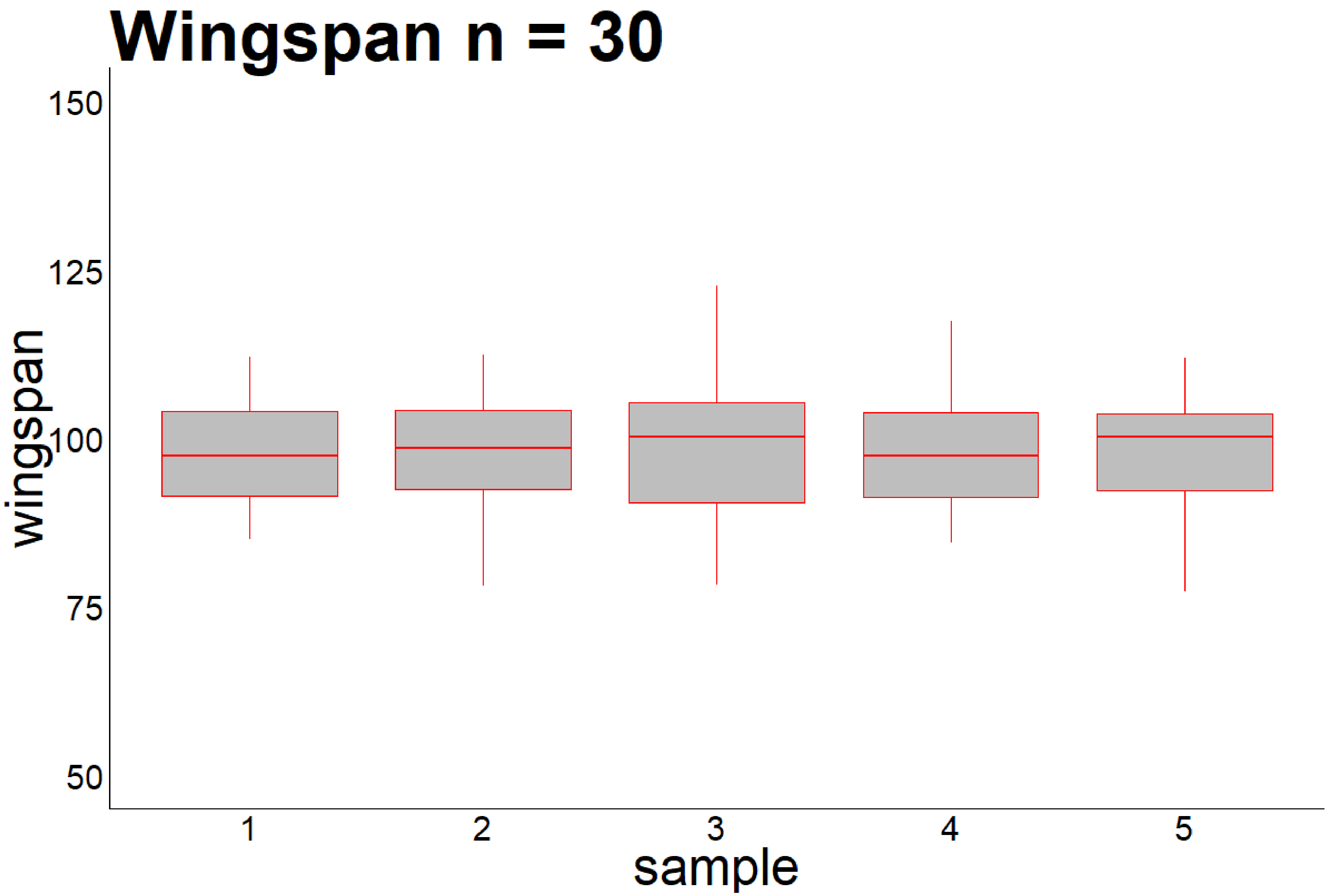

Wingspan n = 30 sample 1
wingspan
150
125
100
75
50
0
10
20
30
Observation

# Wingspan n = 30 sample 5



# Wingspan n = 30

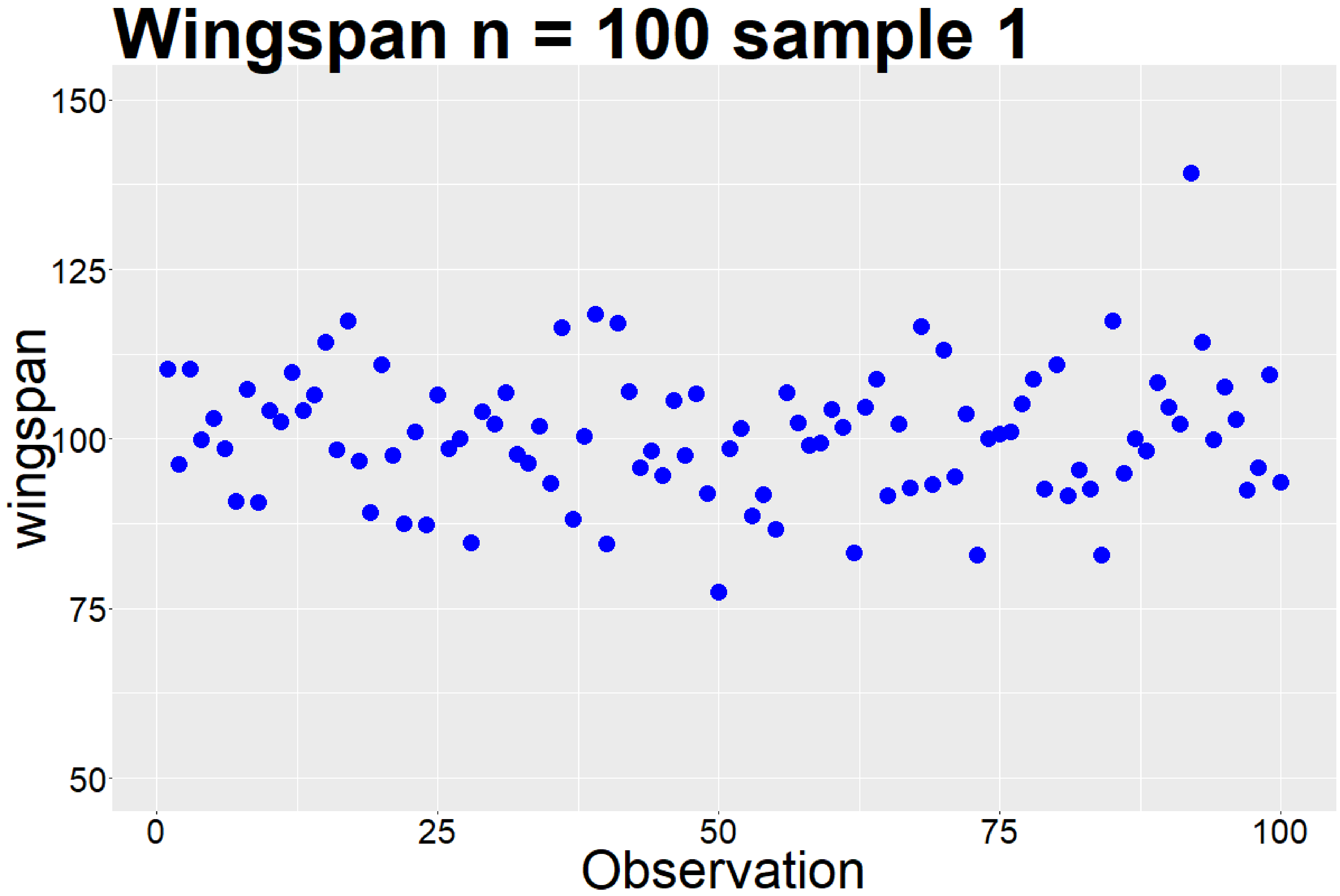



Wingspan n = 30
150
125
100
75
50
wingspan
1
2
3
4
5
sample

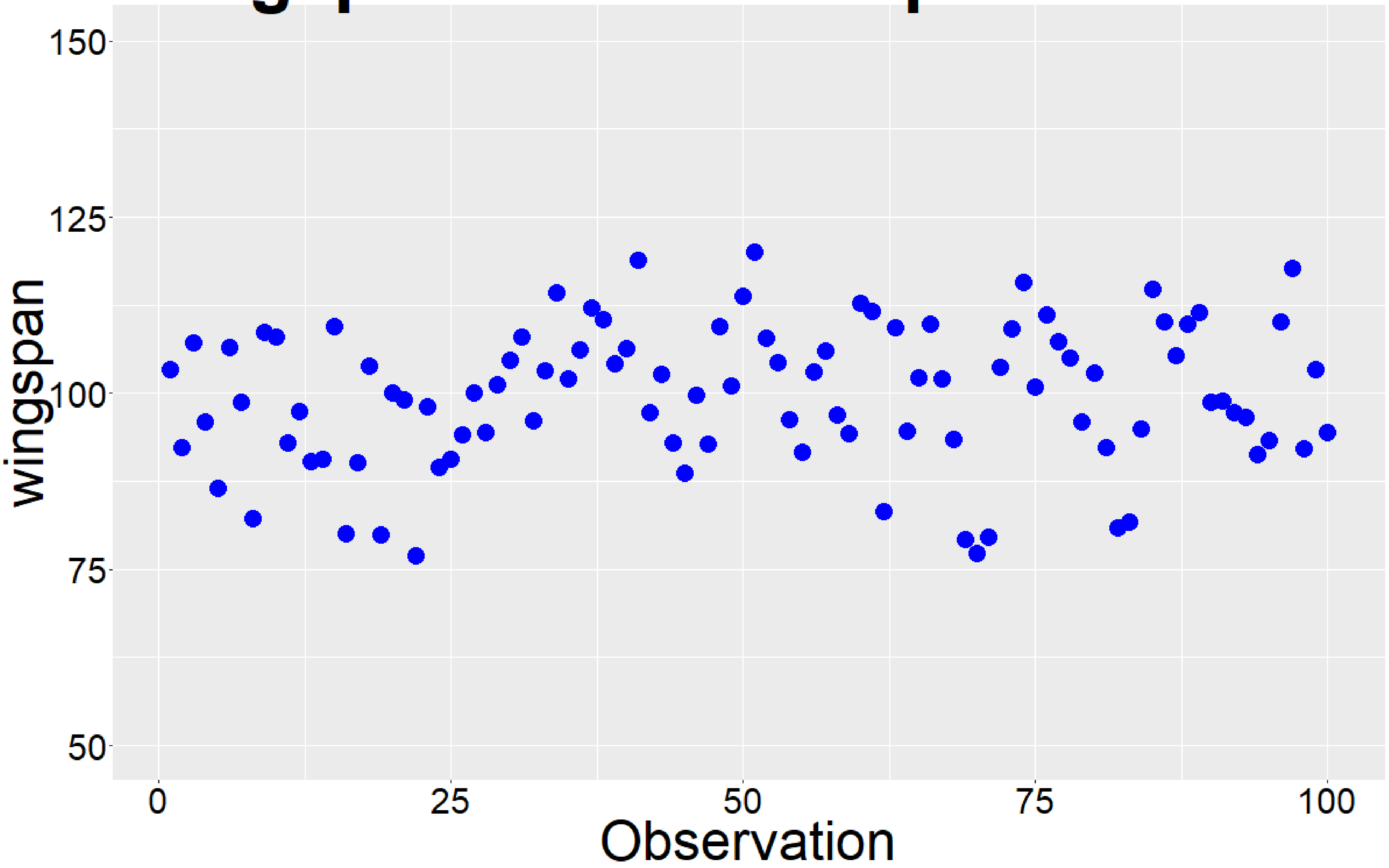
| | Sample | n | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | Whole Population | 10000000 | 100 | 10 | 49.04 | 152.6 |
| 2 | Sample 11 | 30 | 98.12 | 7.498 | 85.02 | 112.2 |
| 3 | Sample 12 | 30 | 97.64 | 9.818 | 78.11 | 112.4 |
| 4 | Sample 13 | 30 | 99.29 | 10.16 | 78.18 | 122.6 |
| 5 | Sample 14 | 30 | 98.48 | 8.655 | 84.55 | 117.4 |
| 6 | Sample 15 | 30 | 98.32 | 8.132 | 77.34 | 112 |

# Sampling: intuition

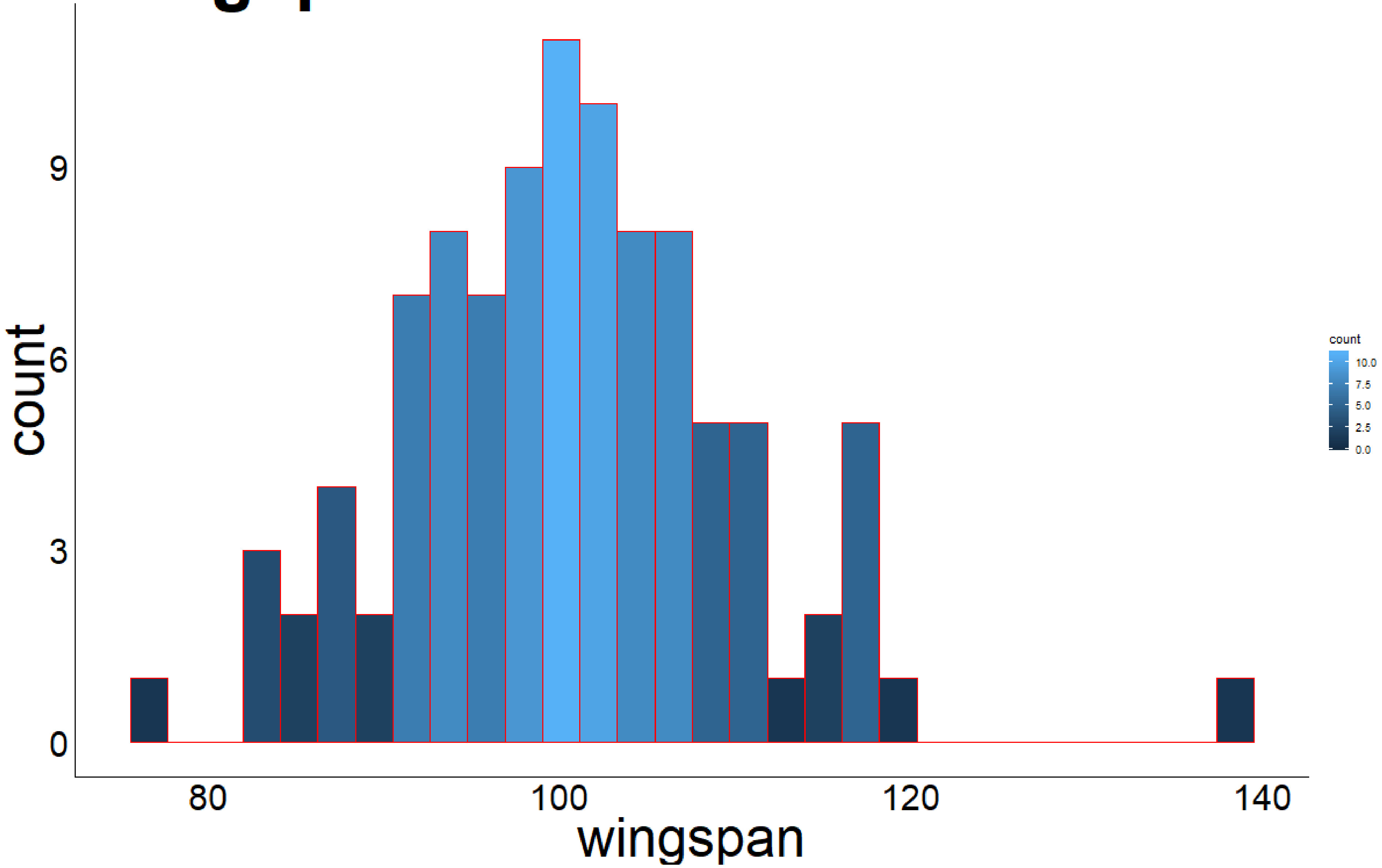
Sample size: $n = 100$

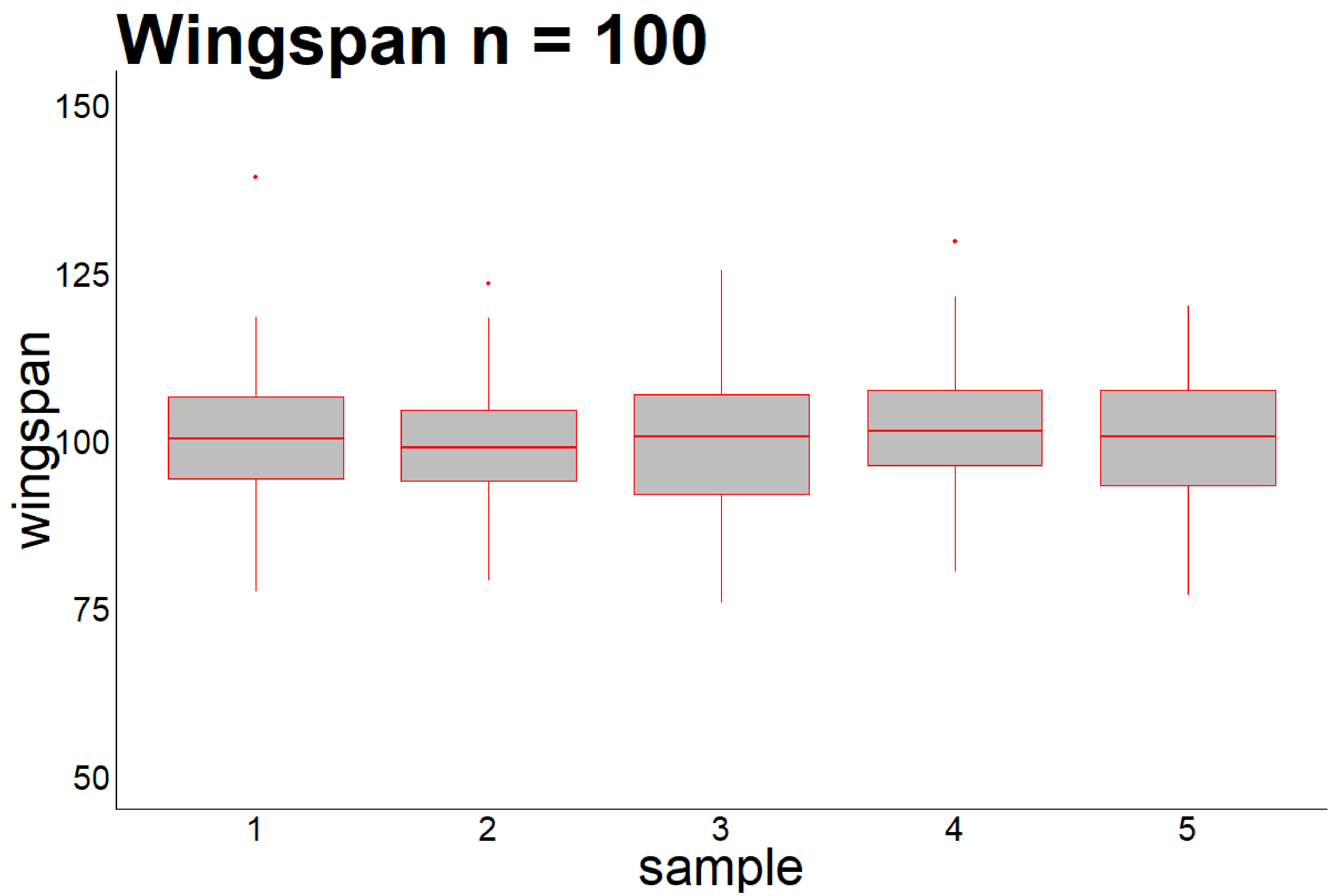

Wingspan n = 100 sample 1
wingspan
150
125
100
75
50
0
25
50
75
100
Observation

# Wingspan n = 100 sample 5



# Wingspan n = 100




Wingspan n = 100
150
125
100
75
50
wingspan
1
2
3
4
5
sample

| | Sample | n | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | Whole Population | 10000000 | 100 | 10 | 49.04 | 152.6 |
| 2 | Sample 16 | 100 | 100.6 | 9.575 | 77.48 | 139.2 |
| 3 | Sample 17 | 100 | 99.49 | 8.122 | 79.08 | 123.3 |
| 4 | Sample 18 | 100 | 99.78 | 9.891 | 75.76 | 125.3 |
| 5 | Sample 19 | 100 | 101.9 | 9.512 | 80.46 | 129.7 |
| 6 | Sample 20 | 100 | 99.77 | 9.993 | 76.92 | 120.1 |

# Variable Scales

1. Scale: definition of ‘scale’ depends on context
2. Scale: measurement system
    1. Intrinsic property of variable, or…
    2. Intrinsic property of how we choose to measure it
        1. E.g. age in years vs. age in age classes

# Discrete and continuous

Discrete: cannot take on intermediate values

1. Counts
2. Categories

Continuous values can assume any value on a continuum*

- * limited by our ability to measure

# Interval and ratio data

Interval: relative zero

1. Degrees C and F
2. Coordinate distance from a fixed point

Ratio: meaningful zero

1. Degrees Kelvin
2. Height, weight, etc

# Data types and analyses

Data type influences mathematical form of analyses:

1. Continuous distributions are often mathematically simpler
2. We can often use continuous distributions to approximate discrete, especially with large sample sizes

# Assignment 1

1. Groups of 4