Statistical Foundations

From theory to practice

Dr. Vianey Leos Barajas + Dr. Sofia Ruiz Suarez Argentina
University of Toronto

Overview of Our Day

- Introduction to random variables + probability distributions
- Sampling mechanisms
- Linear + general linear models
- Including covariates
- Sampling from conditional distributions

- + An Afternoon of Coding in R

Part 1

Random Variables + Their Distributions

Random variables

1st question I ask in interdisciplinary settings is "what is the random variable of interest?"

- A random variable is a function from a sample space S into the real numbers.
- For example, we can think of running an experiment and define random variables of interest as follows:

1 2 3 4 5 6
1 2 3 4 5 6

finite # of outcomes

experiment: toss two dice
random variable: $\boxed{X =}$ sum of the numbers → 2 12

experiment: apply different amounts of fertilizer to corn plants
random variable: X = yield/acre

uncountable
continuous

Random variables

X= the sum of 2 dice

$x = 6$

- The realization of a random variable is the result of applying the function to an observed outcome of a random experiment. It is typically denoted using lowercase italicized Roman letters, e.g. x is a realization of X
- The domain of a random variable is the sample space S, i.e., the set of possible realizations that the random variable can take on.

can it be negative? positive? bounded? countable/ uncountable?

- What is the domain of the random variable X in this case:

X = the outcome
$\{1,6\}, \{1,2\}, \{1,3\}$
$\{6,1\}$

experiment: toss two dice
random variable: X= $\boxed{\text{sum}}$ of the numbers ?

domain = $\{2, 3, 4, \ldots, 12\}$

Types of random variables

$\{1, 2, 3\}$ $\mathbb{N}$ $\{1, 2, 3, \ldots, \ldots\}$

finite infinite

✓

- **Discrete-valued random variable** — domain consists of a countable number of outcomes

  To assign probability mass to the outcomes of a discrete-valued random variable, we construct a probability mass function $\boxed{\text{(pmf)}}$ assigns probability to all possible outcomes

- **Continuous-valued random variable** — domain is uncountable

  To assign probability over an uncountable domain, we construct a probability density function $\boxed{\text{(pdf)}}$ that assign probability over intervals

PMFs/PDFs

$$\sum_x p(X=x) = 1$$

- The probability mass function (pmf) of a discrete random variable X is given by

$$f_X(x) = \underline{P(X = x)} \quad \text{for all } x$$

$$f_X(x) = f(x)$$

assigning a point mass to all possible outcomes

- The probability density function (pdf) of a continuous random variable is a function that satisfies

$$\boxed{\int_a^b f_X(x)} = Pr(a < X < b) \text{ with } \underline{P(X = x) = 0}$$

$$Pr(a \le X \le b)$$

$$Pr(X = a) = 0$$

Expectations and Variances

- The expectation (i.e. mean) and variation around the mean (i.e. variance) are common quantities of interest in statistics. We use the pmf/pdf to compute them.

- Let $E(X)$ be the expectation a random variable X, then $E(X) = \int xf(x)dx$ in the continuous case and $E(X) = \sum xf(x)$ in the discrete case where $f(x)$ represents the probability mass/density function

- The variance is $E(X^2) - [E(X)]^2$

$E(x^2) = \int x^2 f(x)\, dx$

Discrete-valued random variables and common probability mass functions

Bernoulli

- domain $\{0, 1\}$
- assign probability mass to all outcomes / 0 + 1

- The Bernoulli distribution is a probability mass function that takes the values 0 and 1 only.

- If we let the outcome of a coin flip be only heads or tails, we can assign heads as 0 and tails as 1, and construct a probability mass function in the following manner:

$P(X=0) + P(X=1) = 1$

$\cancel{p+p} + P(X=1) = 1-p$

$$P(X=0) = \boxed{p} \text{ and } P(X=1) = 1-p \text{ with the assumption that } p \in (0,1)$$

- We often denote the distribution with *parameter* p as $Bern(p)$. It has expectation p and variance $p(1-p)$.

Binomial — counting

$E(Y) = \sum y \binom{N}{y} p^y (1-p)^{N-y}$

domain $\{0, 1, 2, \ldots, N\}$

- The binomial distribution is connected to the Bernoulli distribution. Let X_i be identically distributed Bernoulli random variables where our interest is in $\sum_{i=1}^{N} X_i = Y$. That is, out of N trials, what is the probability of y successes?

each trial you ~~can~~ get a 0 or 1

- The binomial distribution has pmf $\binom{N}{y} p^y (1-p)^{N-y}$ with expectation Np and variance $Np(1-p)$

$N = 5 \quad p = 0.4$

$Pr(Y=2) = \binom{5}{2} \cdot (0.4)^2 (1-0.4)^{5-2}$

- We typically write it as Binomial(N, p)

fixed

Poisson —

Bernoulli $\{0, 1\}$

Binomial $\{0, \ldots, N\}$

often used in contexts where the random variable is a count

- The Poisson distribution is a distribution for data that takes on values 0, 1, 2, 3,… ∞ and so on.

- It has the property that the expectation is equal to its variance, with λ reflecting these quantities

- It has the form:

$$\frac{e^{-\lambda}\lambda^x}{x!} \text{ for } \lambda > 0$$

Pois(λ)

Pois$(\lambda = 2.2)$

$$\Pr(X = 5) = \frac{e^{-2.2}(2.2)^5}{5!}$$

domain $\{0, 1, 2, \ldots, \infty\}$

Negative Binomial (one of its parametrizations)

domain $\{0, 1, 2, \ldots, \infty\}$

- An alternative manner to assign probability mass to a random variable with domain 0, 1, …. ∞ Is to use the negative binomial distribution.

- It has the following form:

$$\boxed{\binom{x+\phi-1}{x}\left(\frac{\mu}{\mu+\phi}\right)^{x}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}}$$

scientist often question whether to use the poisson or neg. binomial dist

- With ~~exception~~ μ and variance $\boxed{\mu + \frac{\mu^2}{\phi}}$

expectation

$NB(\mu, \phi)$

$NB(\mu=1, \phi=2)$

$Pr(X=5) \quad \binom{5+2-1}{5}(\ldots)(\ldots)$

Discrete-valued distributions

Comparisons

- Bernoulli
- Binomial
- Poisson
- Negative Binomial

Continuous-valued random variables and common probability density functions

Normal/Gaussian

- The normal/Gaussian distribution is one of the most commonly used distributions — the most common perhaps? It is defined over the real line.

$\rightarrow \mathbb{R}$

- It has the form:

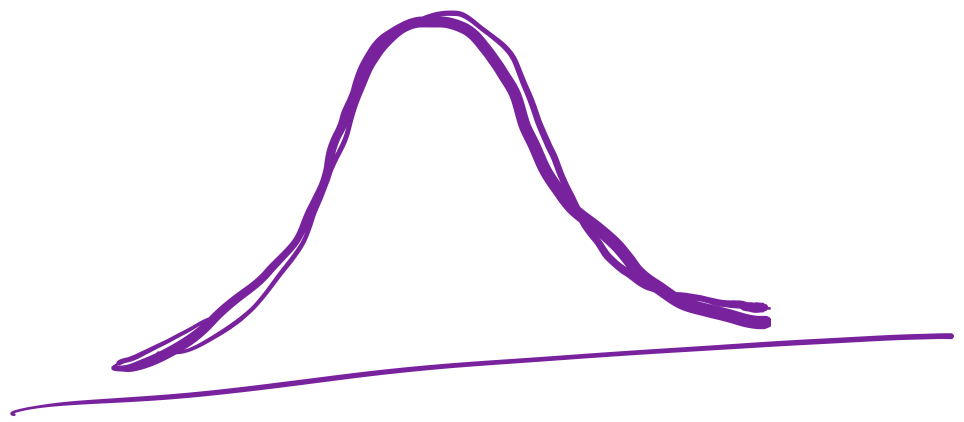
$$\frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

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

mean

variance

- With expectation μ and variance σ^2



Gamma

— one parametrization
domain $(0,\infty)$

— often used to describe the speed at which sharks swim

- Continuous-valued random variables can be defined over only some parts of the real line. For example, it's quite common to have $X > 0$ if we think of quantities that must be measured but can not be negative. One common distribution that is used is the gamma distribution as it only assign probability mass on the positive real line.

- It has the form:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x) \text{ with expectation } \alpha/\beta \text{ and variance } \alpha/\beta^2$$

Continuous-valued distributions

Comparisons

- Normal

- Gamma

Simulation

an effective way to communicate across disciplines and [also gain intuition about statistical models]

Simulation

Main points:

- Learn about a statistical model by the data it generates

- Think about the data generated via different sample sizes

- Be able to compare simulated data from different models and across sample sizes

Standard Normal Distribution

$\mu \in \mathbb{R}$

$\sigma^2 > 0$

distributed as a

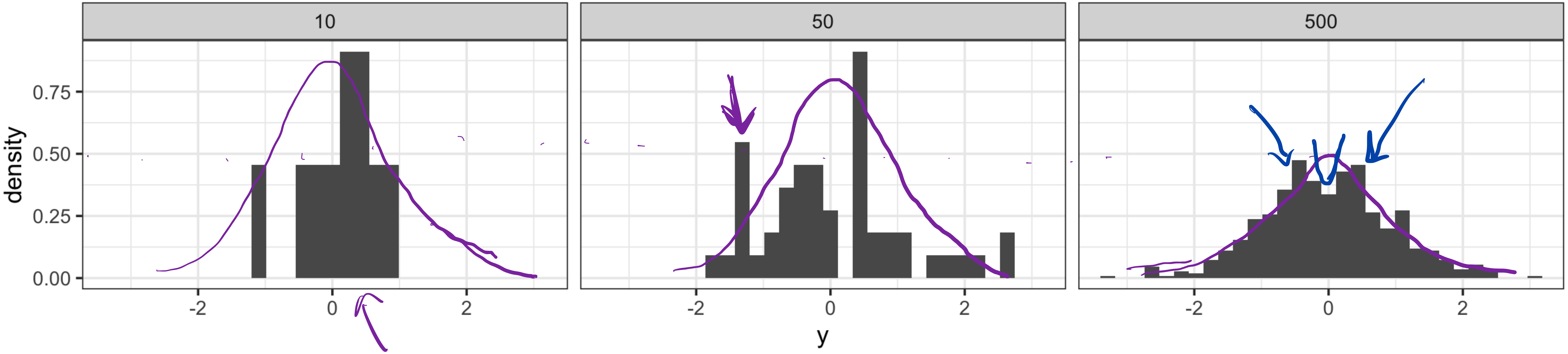
$$y_i \sim N(\mu, \sigma^2) \quad \textit{for} \quad i = 1,2,\ldots,I$$

normal distribution with mean μ and variance σ^2

sample size I

standard deviation

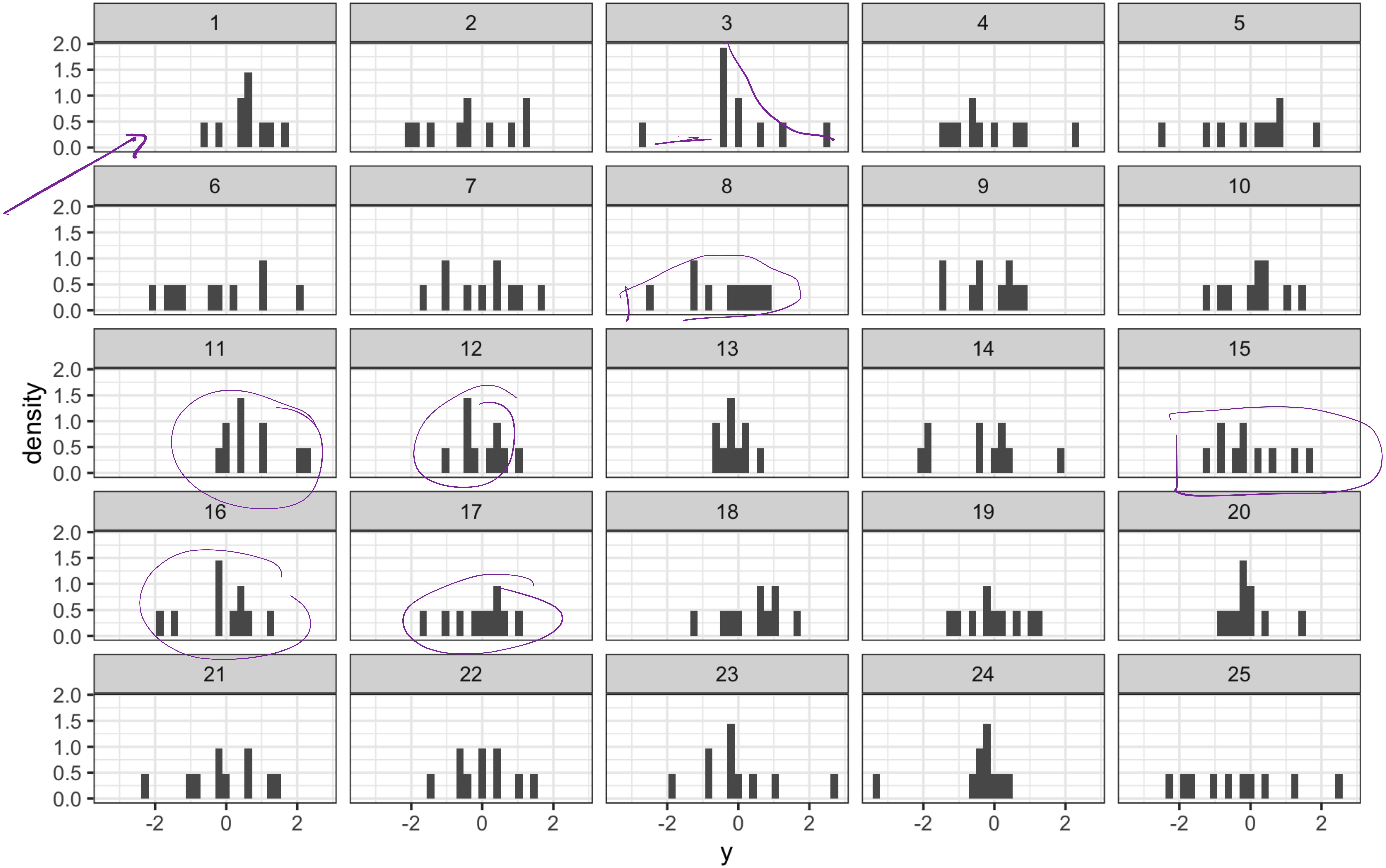
$I = 10, \mu = 0, \sigma = 1$	$I = 50, \mu = 0, \sigma = 1$	$I = 500, \mu = 0, \sigma = 1$



25 Replicates

$I = 10, \mu = 0, \sigma = 1$

How do these compare?

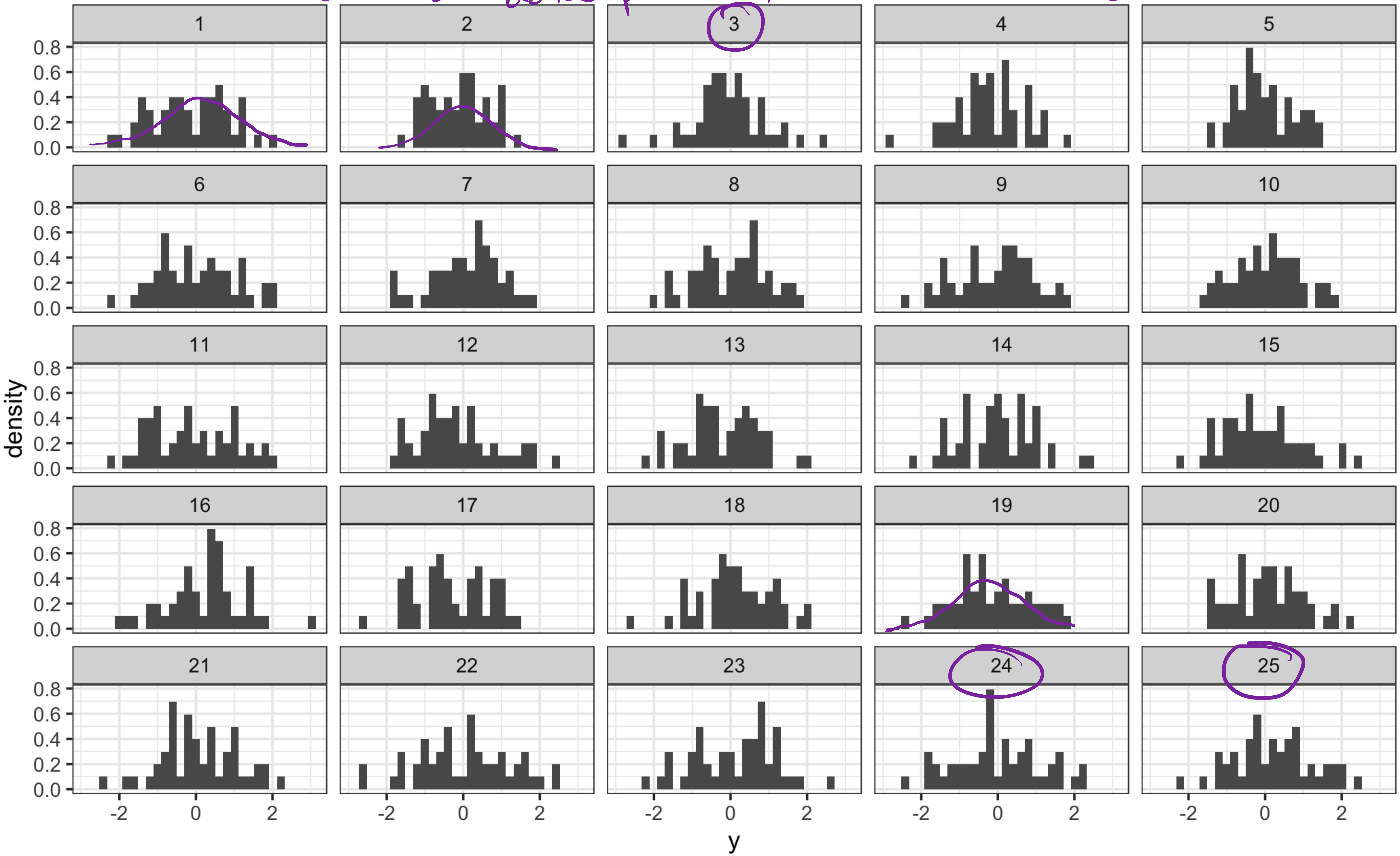


25 Replicates

$I = 50, \mu = 0, \sigma = 1$

How do these compare?

what does 50 data points from a normal dist look like?

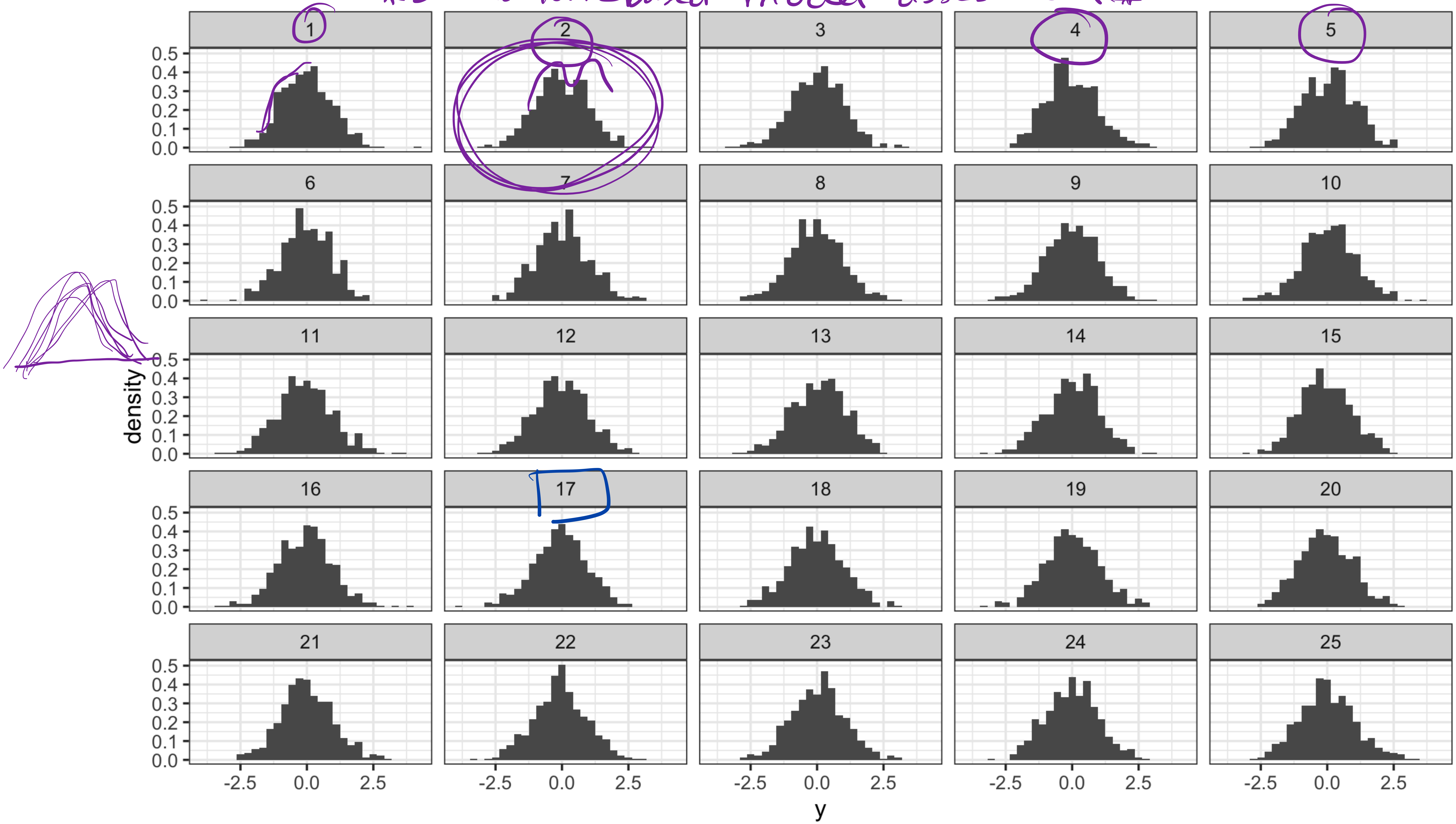


25 Replicates

$I = 500,\ \mu = 0,\ \sigma = 1$

How do these compare?

simulation-based model assessment



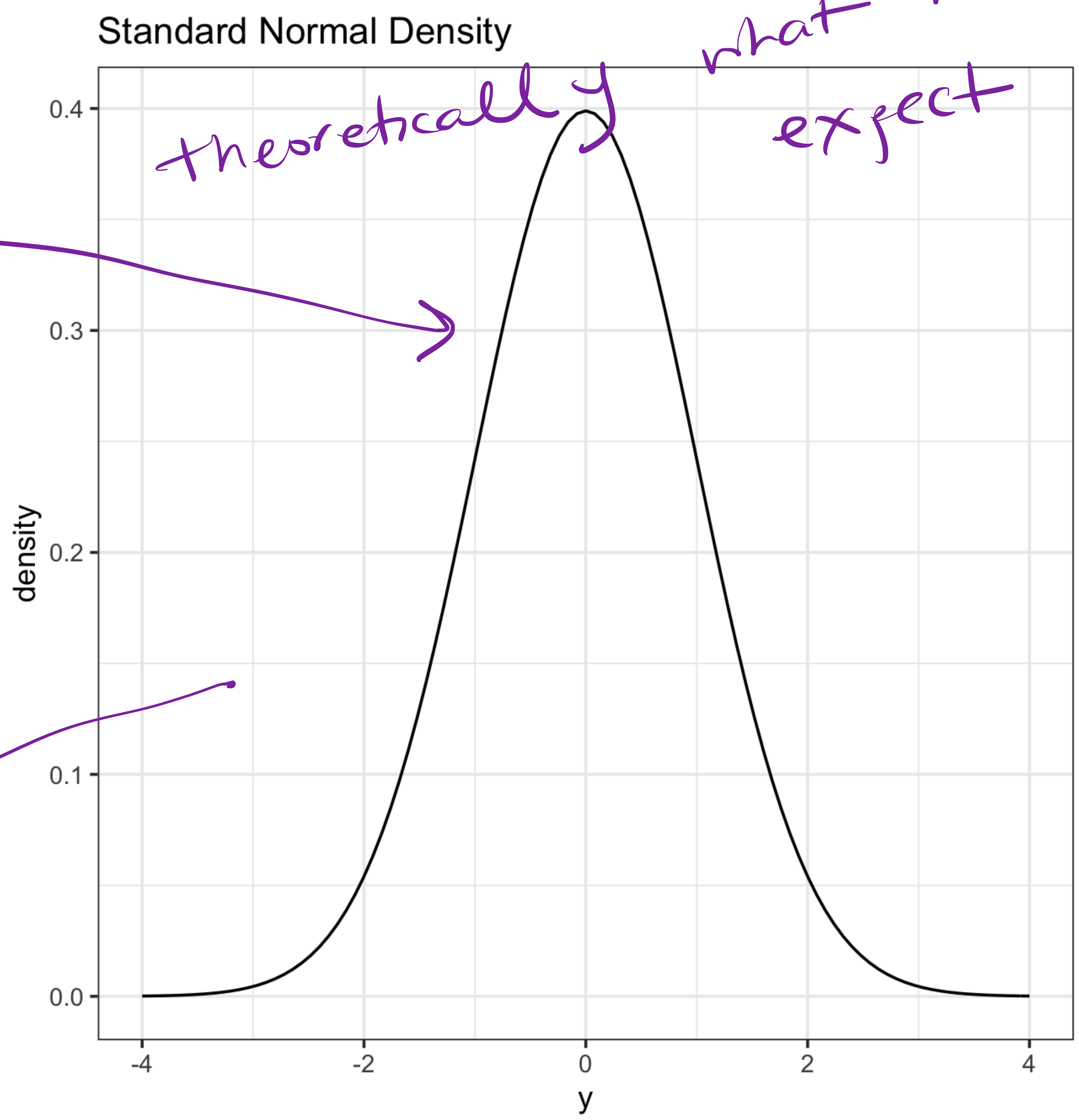
Standard Normal Distribution

- What's the point of comparing replicates of various sample sizes from a standard normal distribution?

- How does sample size affect how 'similar' the draws are?

- We know the 'true model' — compare the mathematical description vs the data generated.

theoretically what we expect

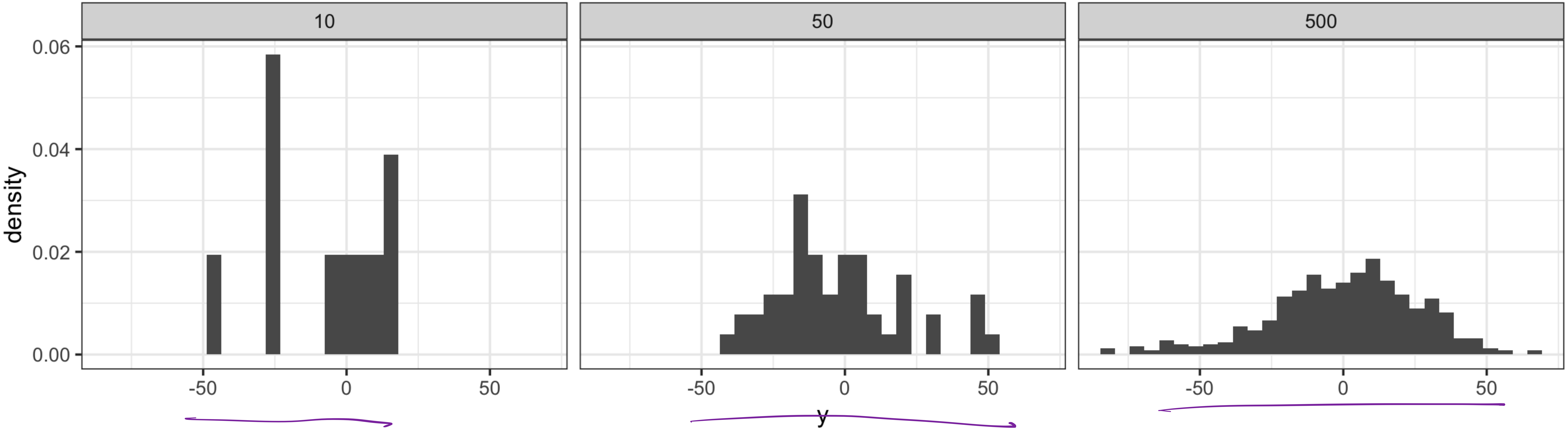


Normal Distribution with Larger Variance

$$y_i \sim N(\mu, \sigma^2) \quad for \quad i = 1,2,\ldots,I$$

$\sigma^2 = 25^2$

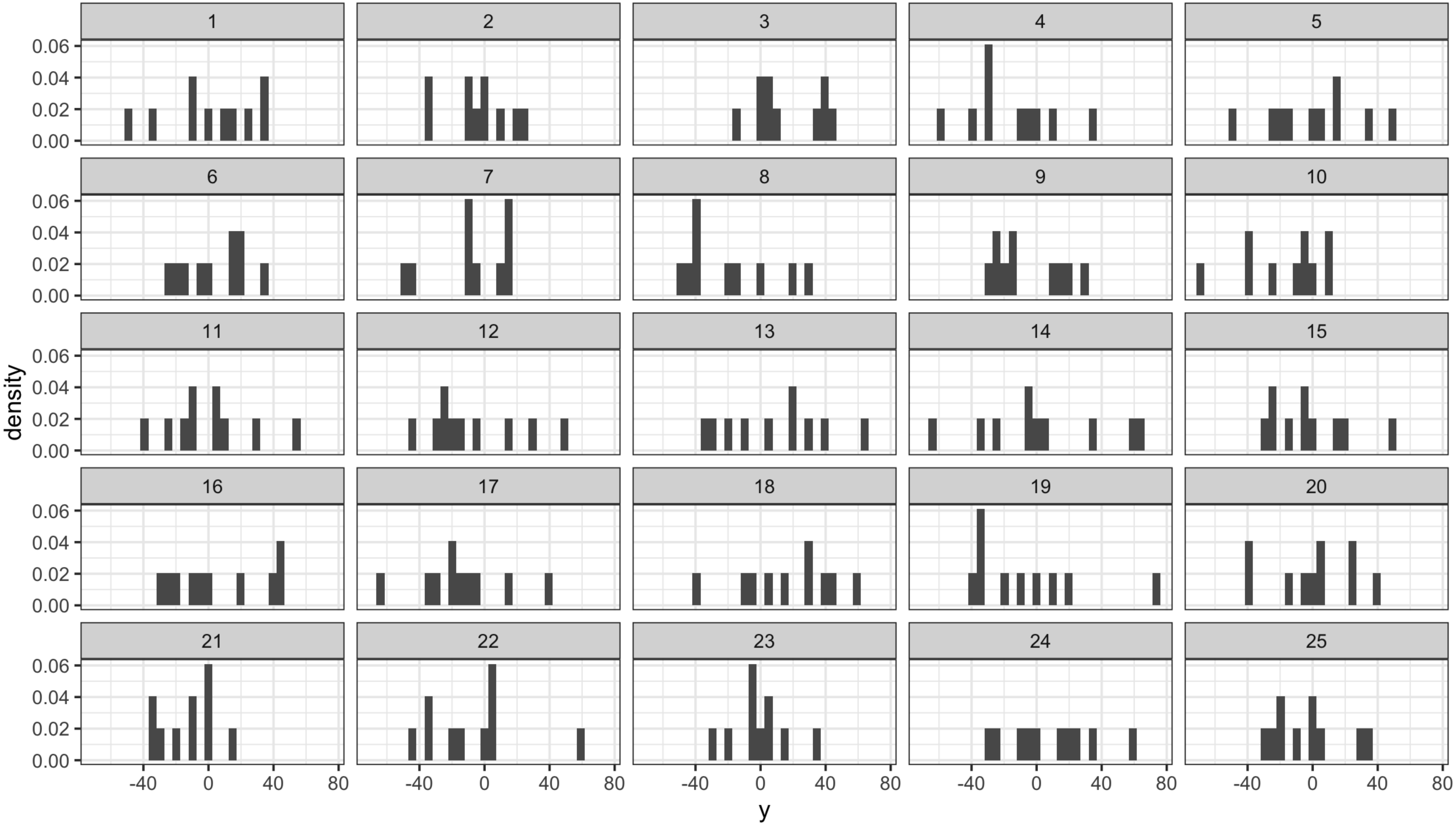
$I = 10,\ \mu = 0,\ \sigma = 25$	$I = 50,\ \mu = 0,\ \sigma = 25$	$I = 500,\ \mu = 0,\ \sigma = 25$



25 Replicates

$I = 10, \mu = 0, \sigma = 25$

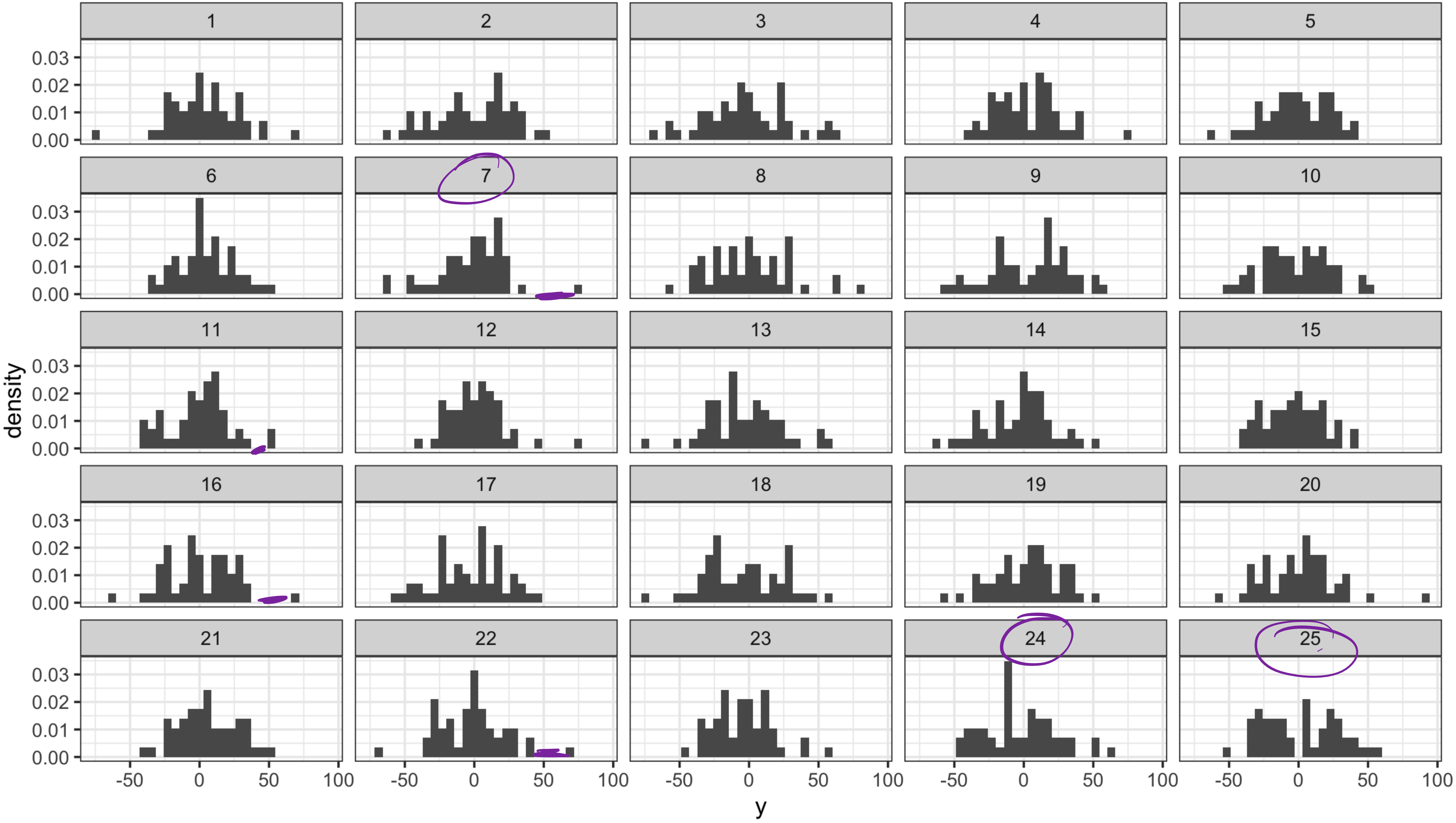
How do these compare?



25 Replicates

$I = 50, \mu = 0, \sigma = 25$

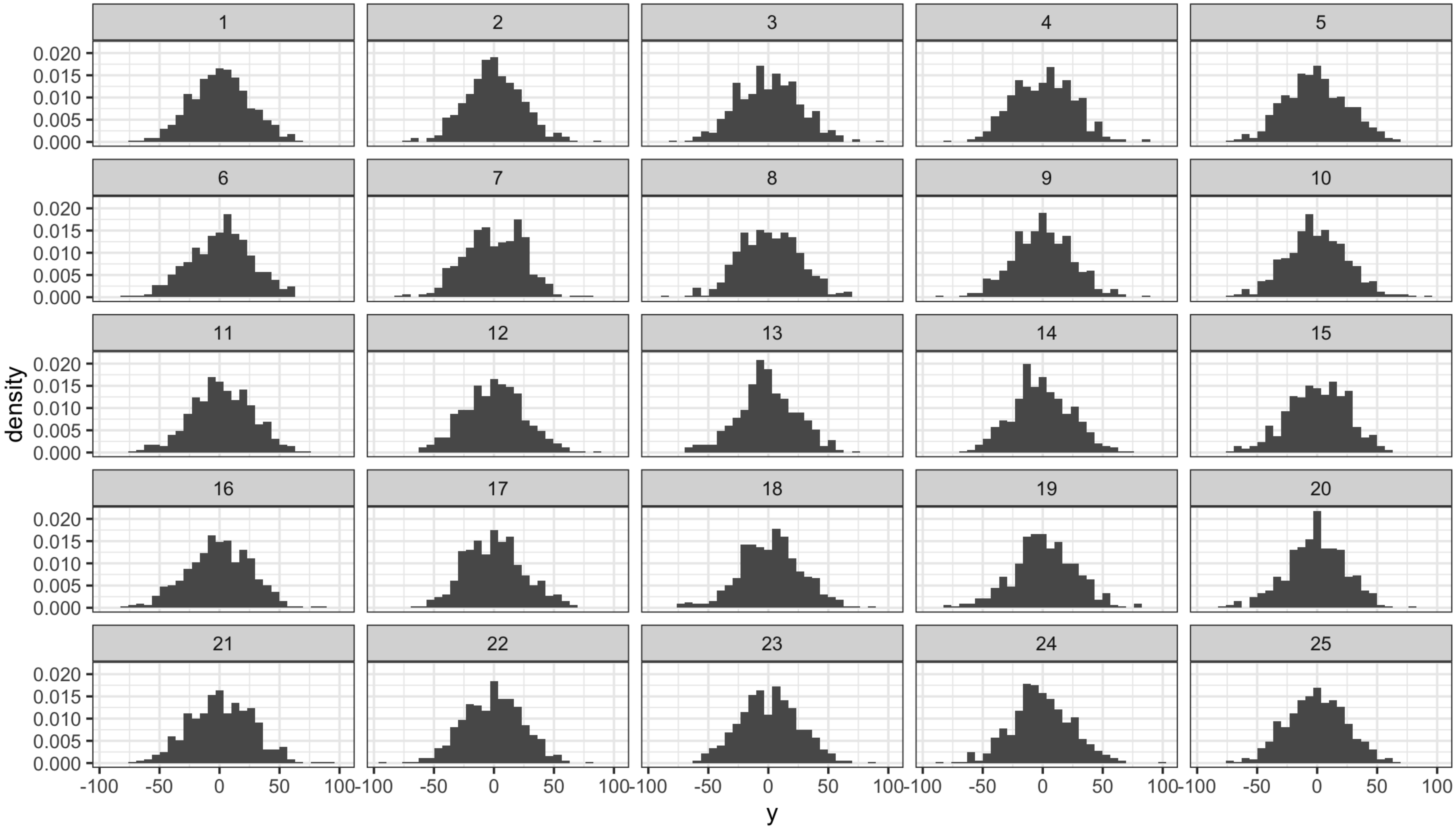
How do these compare?



25 Replicates

$I = 500, \mu = 0, \sigma = 25$

How do these compare?

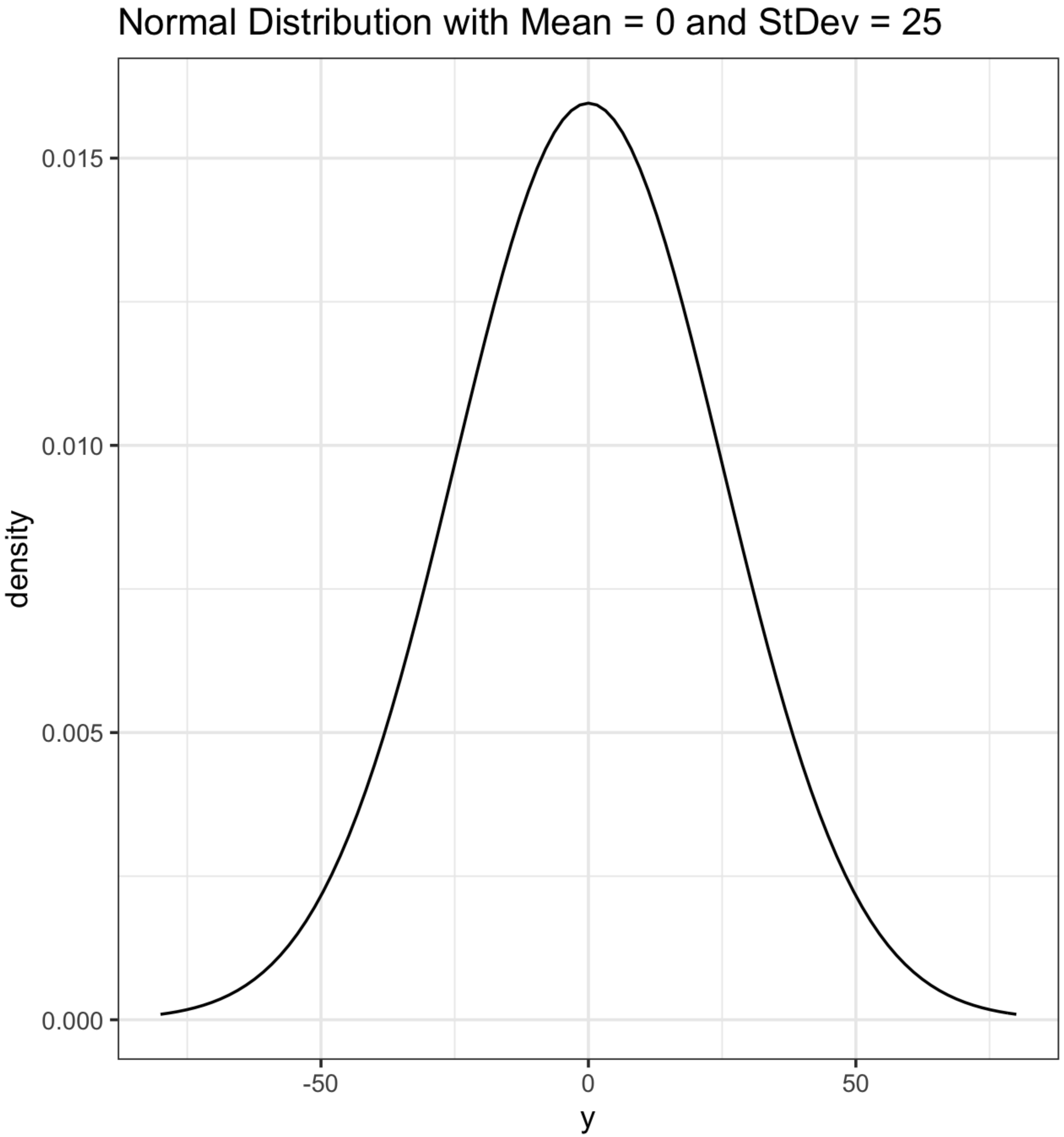


Normal Distribution Again…

- What's the point of comparing replicates of various sample sizes from a standard normal distribution?

- How does sample size affect how 'similar' the draws are?

- We know the 'true model' — compare the mathematical description vs the data generated.



Quick Reflection

What does it mean to connect data to a statistical model?

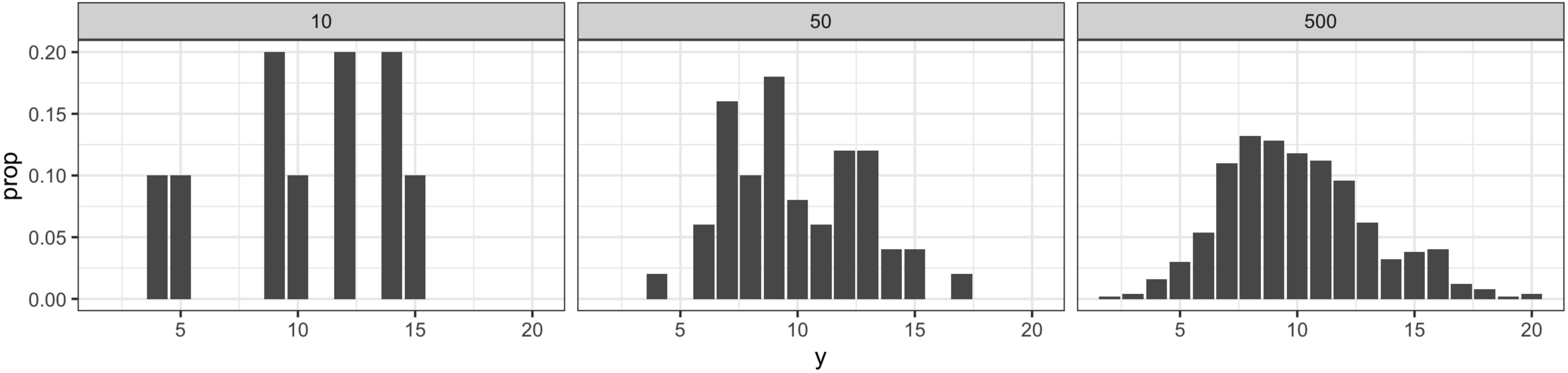
Should we compare our data to ***all possible data sets of the same sample size our model *could have generated**** ?

Poisson Distribution

expectation and variance is equal to 10

$$y_i \sim Poisson(\lambda) \quad for \quad i = 1,2,\ldots,I$$

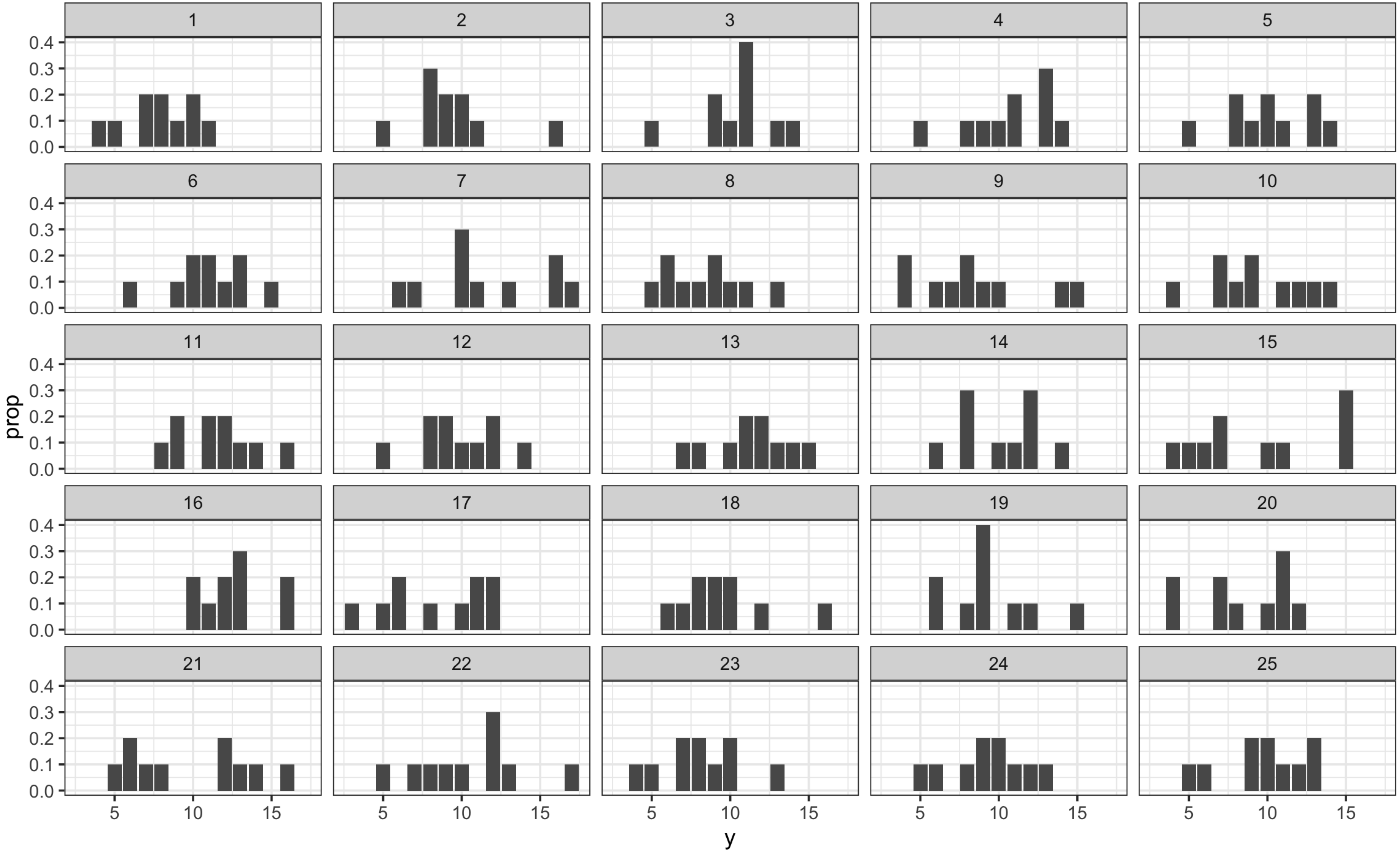
$I = 10, \lambda = 10$	$I = 50, \lambda = 10$	$I = 500, \lambda = 10$



25 Replicates

$I = 10, \lambda = 10$

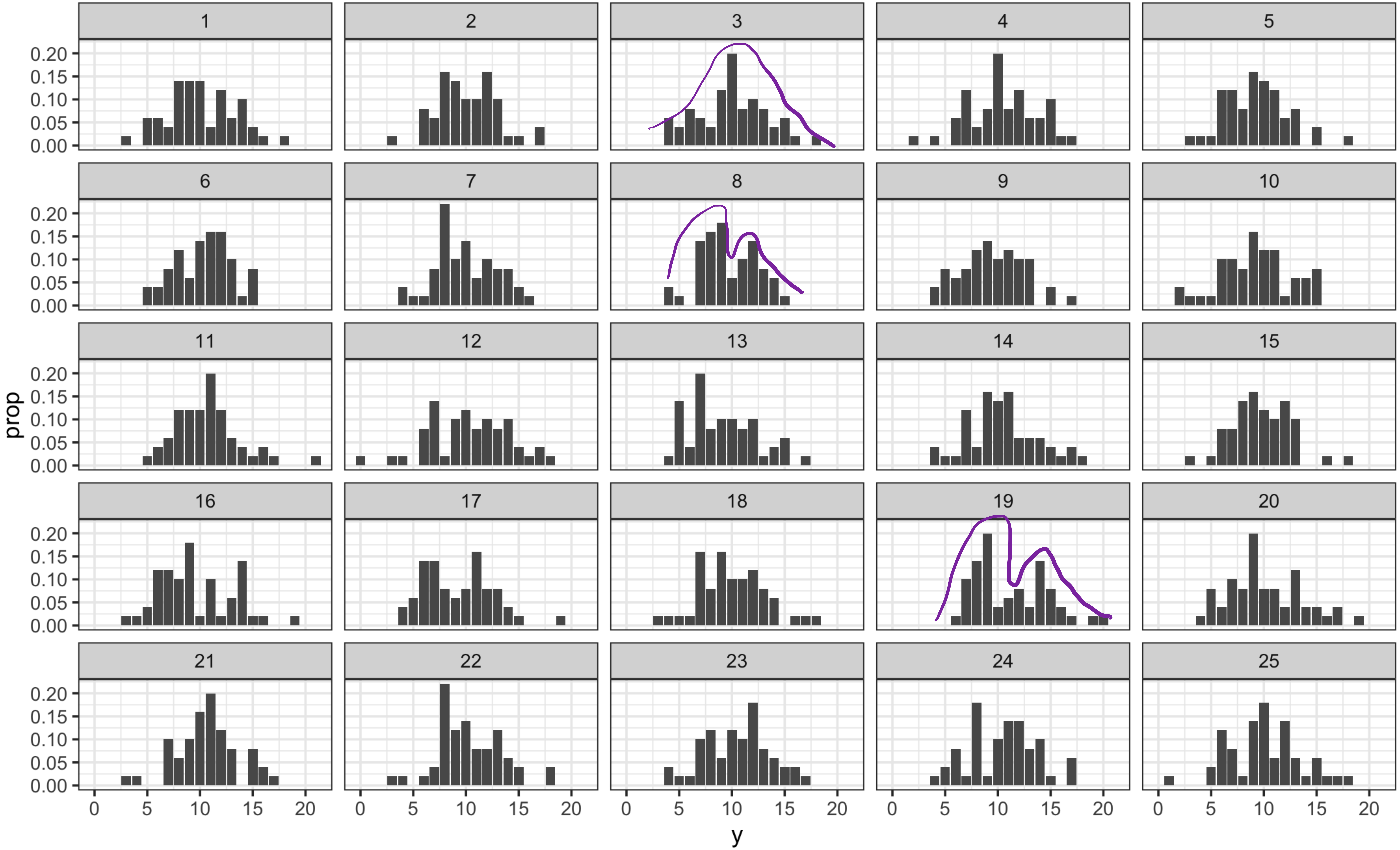
How do these compare?



25 Replicates

$I = 50, \cancel{\mu = 0, \sigma = 25}$ $\lambda = 10$

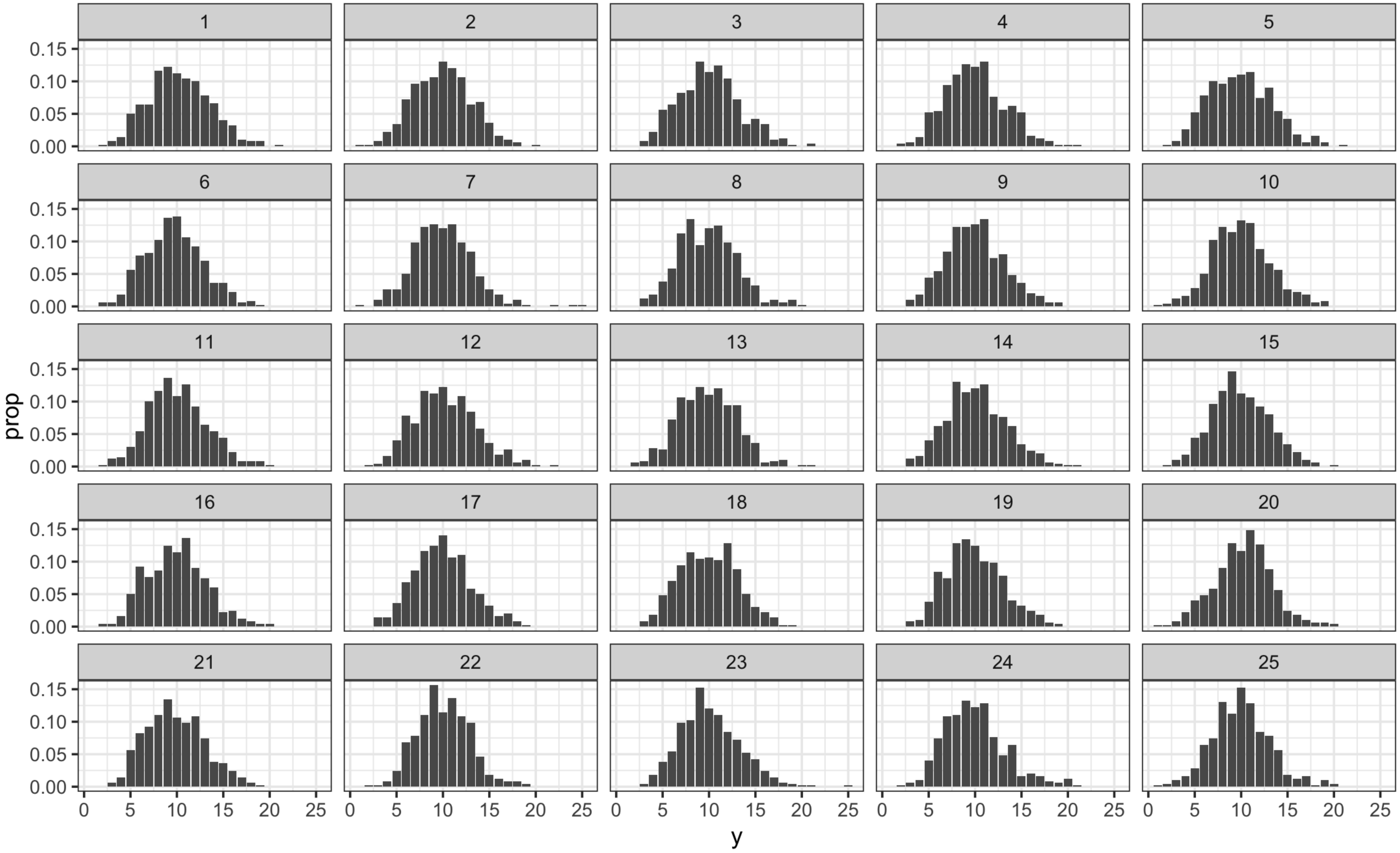
How do these compare?



25 Replicates

$I = 500, \cancel{\mu = 0, \sigma = 25}$ $\lambda = 10$

How do these compare?



How else to compare across samples?

- We can compute sample means and standard deviations: $\bar{x}, \ s$
- We can compute sample quantiles
- Depending on what you're trying to represent….think of the simulated data as actual data that was collected. Does it make sense?

This is my secret weapon when working on interdisciplinary projects. It can be hard to describe a mathematical object to folks without quantitative training, but **everyone understands data**.

Any questions?

Part 2

Simple linear regression (SLR)

Simple Linear Regression

extends the normal dist

observation covariate line

Often written as: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad for \quad i = 1,2,\ldots,I$

independently + identically distributed

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$\mu = 0,$

But really, it all goes back to probability distributions:

given / conditional on

$$y_i \,|\, x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

distributed as

Specifically, we're now in the realm of conditional distributions.

Conditional Distributions Run the World

- Conditional distributions allow us to characterize distributions of a random variable under a fixed value x_i, e.g. $f(y \mid x_i) \neq f(y \mid x_j)$ if $x_i \neq x_j$

  outcome of a die roll differs under different temperatures

- This is still regression. But it's good to understand that under all of the lines and curves we fit, probability mass functions and distribution functions are the real stars.

Steps to simulate

1. Suppose we have values of x_i — fixed

2. Fix values of β_0 and β_1

3. Fix a value of σ^2

4. Compute values of $\beta_0 + \beta_1 x_i$

5. Simulate I values from $N(0,\sigma^2)$ using `rnorm' function in R

6. Add values from step 4. and step 5. to obtain simulations of y_i

deterministic

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \epsilon_i \quad for \quad i = 1,2,\ldots,I$$

$$\boxed{\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)}$$ — noise

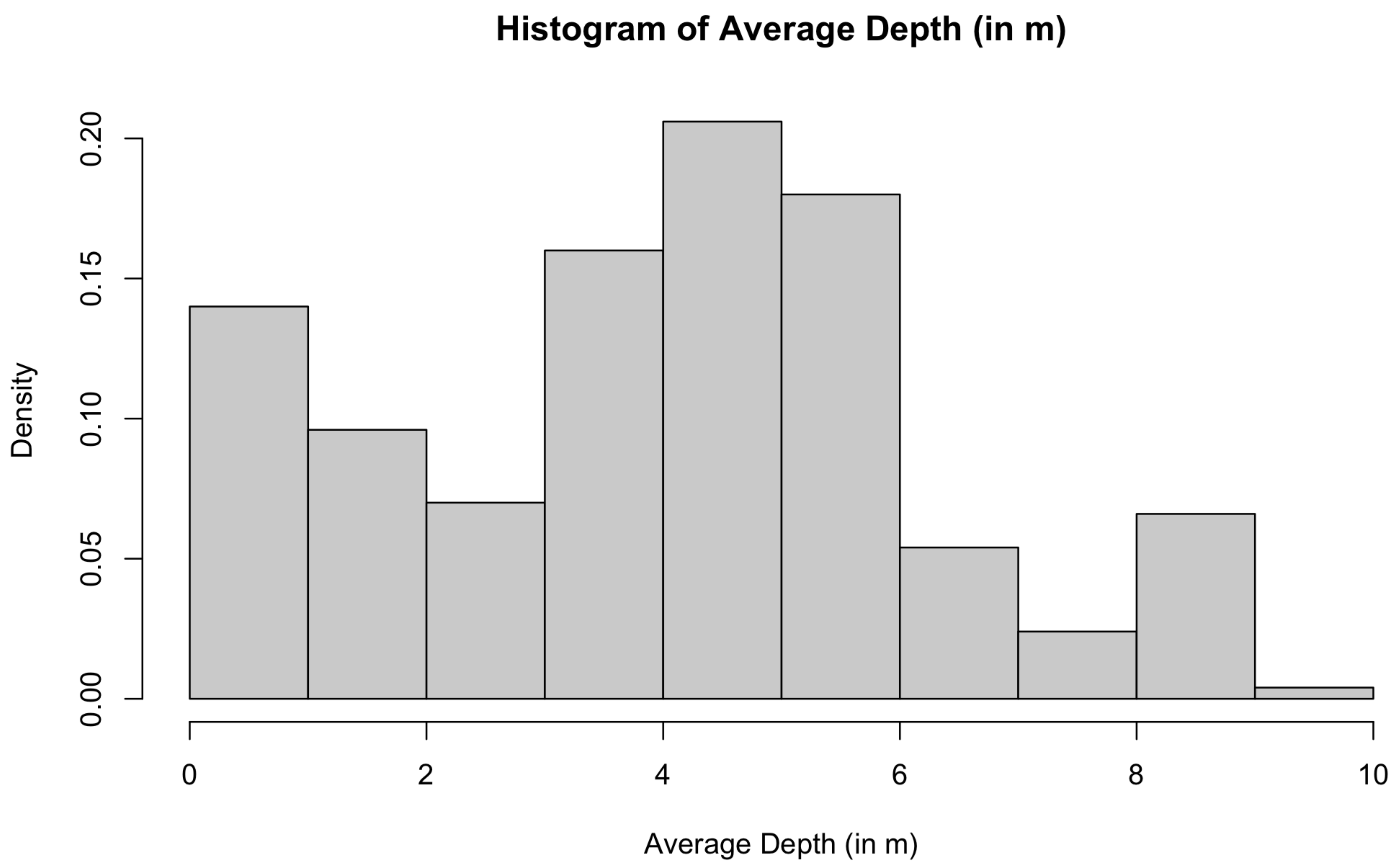
stochastic

Some key takeaways:

- For a set of specific values of $\{\beta_0, \beta_1, \sigma\}$, we generate a set of I values of y_i
- With the same values of $\{\beta_0, \beta_1, \sigma\}$, we can generate many replicates of y_i, each of sample size I
- J^{th} replicate: $\{y_{1,j}, y_{2,j}, \ldots, y_{I,j}\}$

covariate

Let's take x_i for $i = 1,\ldots,500$ to be average depth (at 1 sec intervals) of a shark



Example 1:

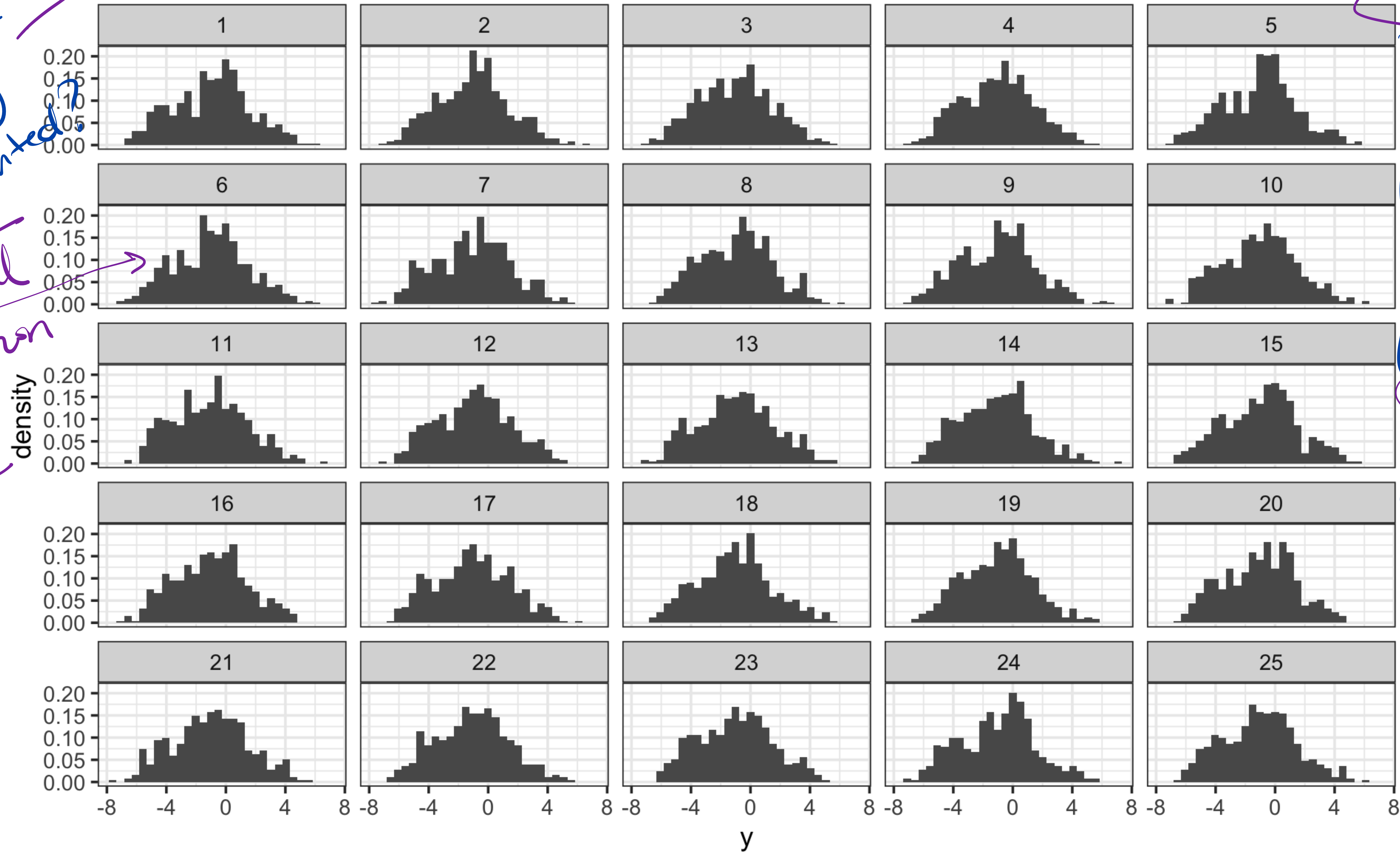
$\{\beta_0 = -5, \beta_1 = 1, \sigma = 1\}$

$\{Y_{1j}, Y_{2j}, \ldots, Y_{Ij}\}$

cond. dist

Is Y_i normally distributed? — not necessarily

marginal distribution of Y



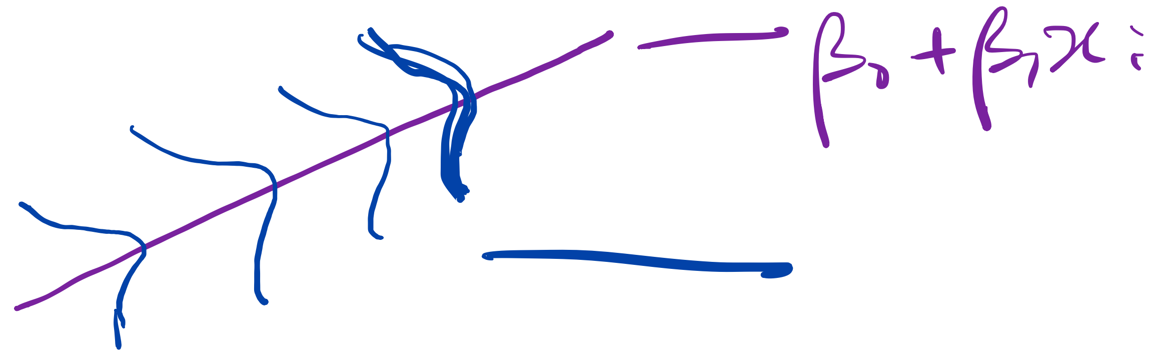
$Y_i | X_i$ is a normal dist

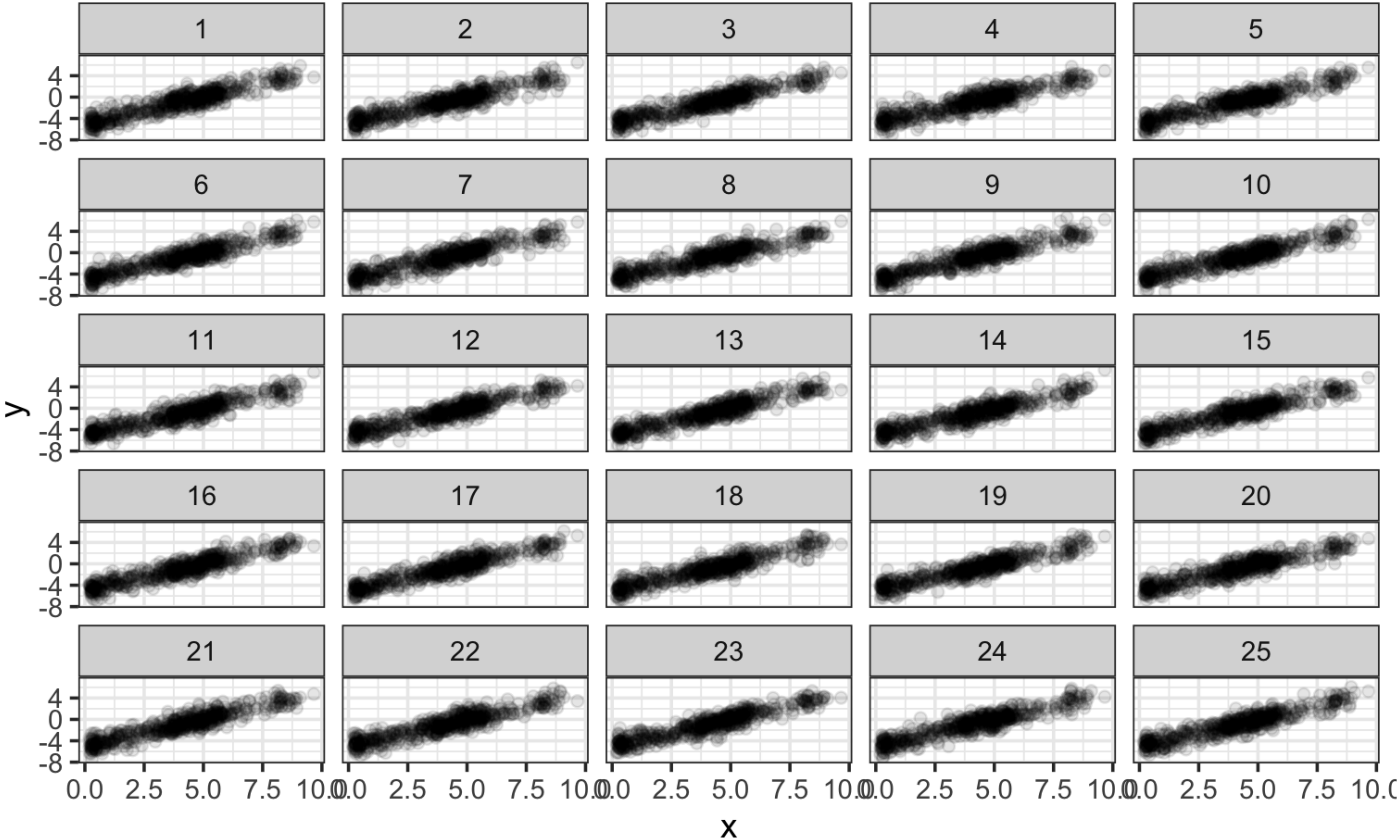
marginal ~~the~~ dist

dist of Y_i might not be

Example 1:

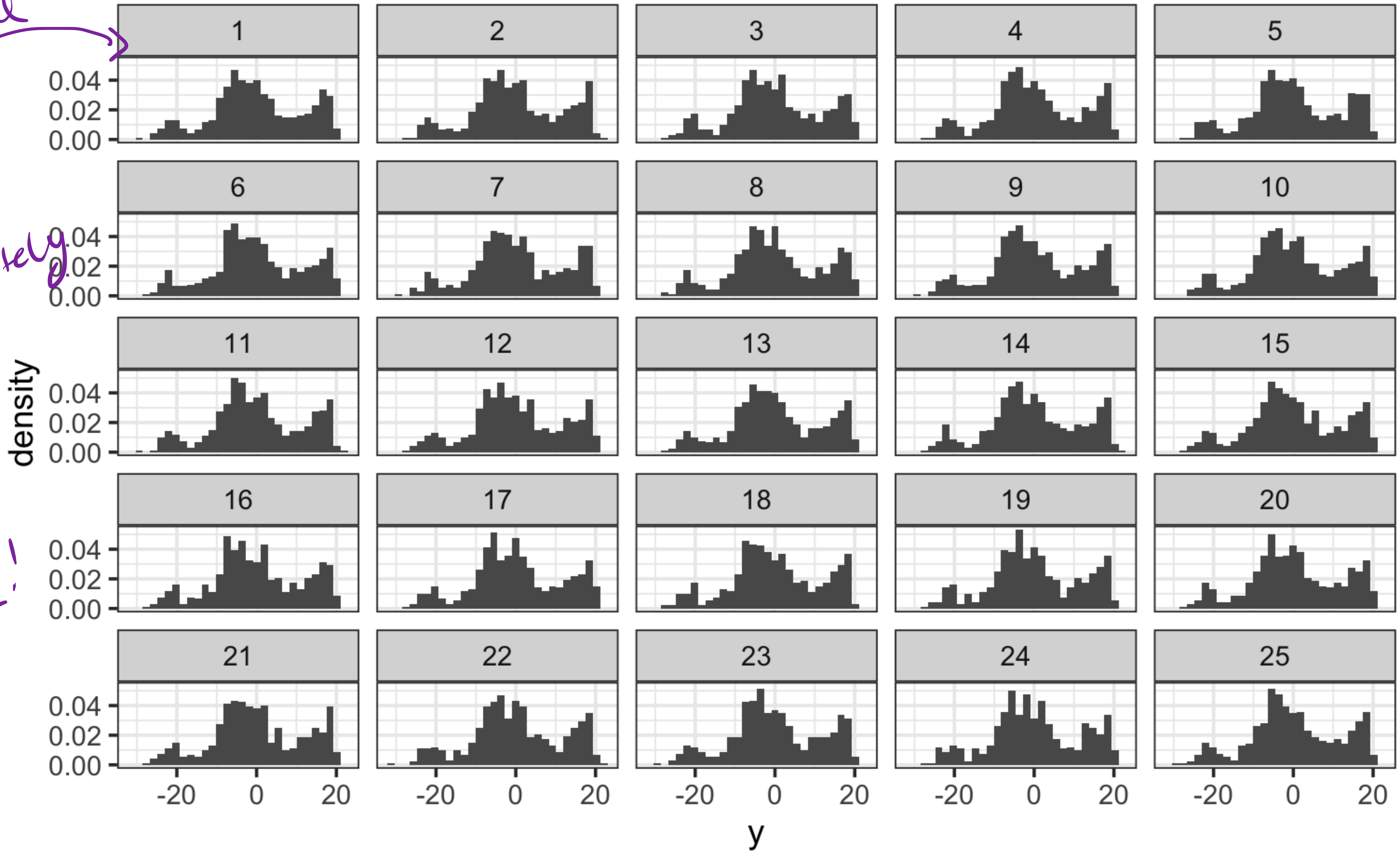
$$\{\beta_0 = -5, \beta_1 = 1, \sigma = 1\}$$





Example 2: $\{\beta_0 = 20, \beta_1 = -5, \sigma = 1\}$

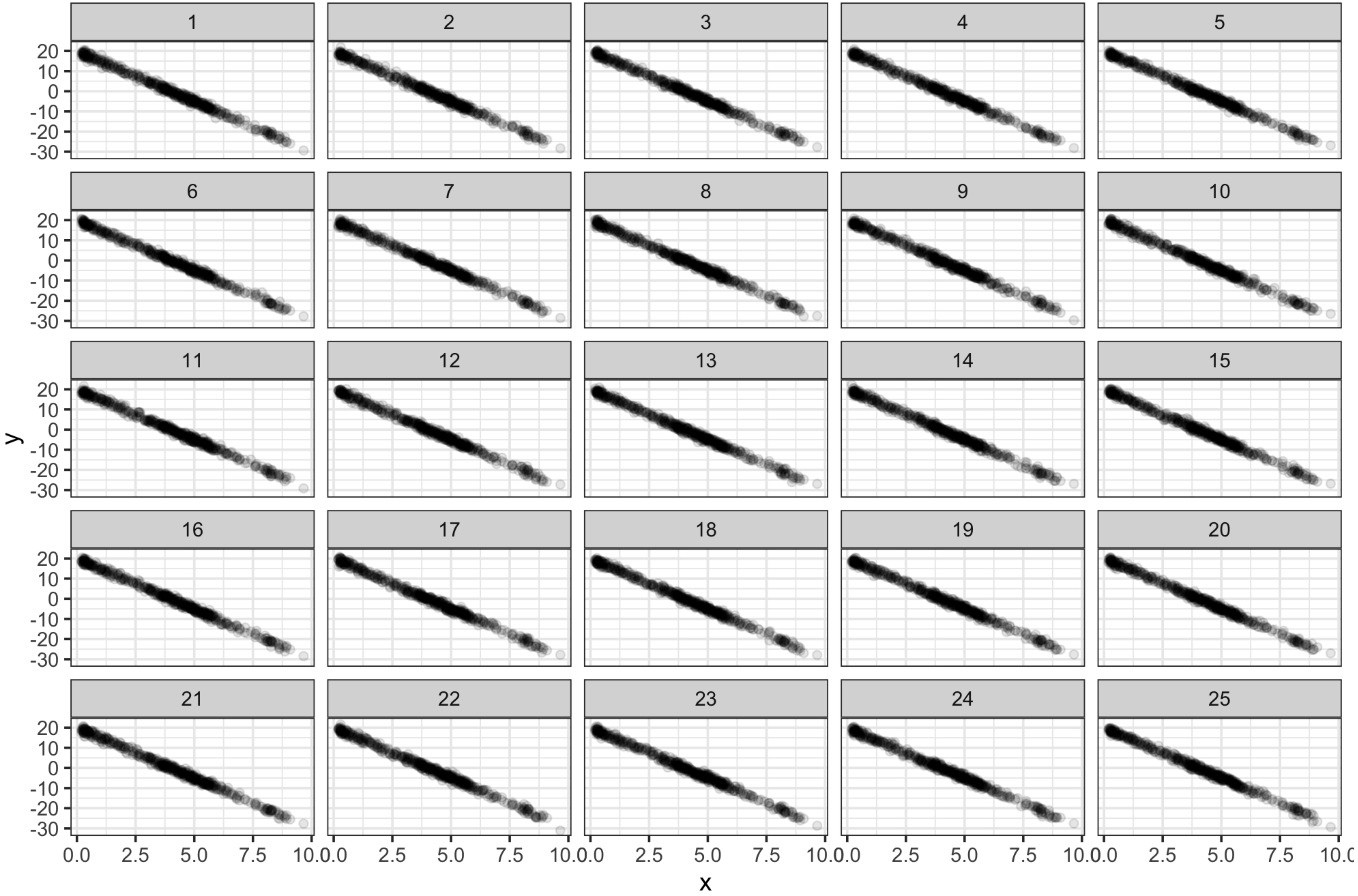
marginal dist of Y which is definitely not normal AND that's okay!



Example 2:

intercept slope

$$\{\beta_0 = 20, \beta_1 = -5, \boxed{\sigma = 1}\}$$



What do we see so far?

- Let's compare the value of β_1 to σ across the two examples— does it make a difference?

  $\beta_1 = -5$ or 1, $\beta_1 = -5$ the points around the line look more condensed

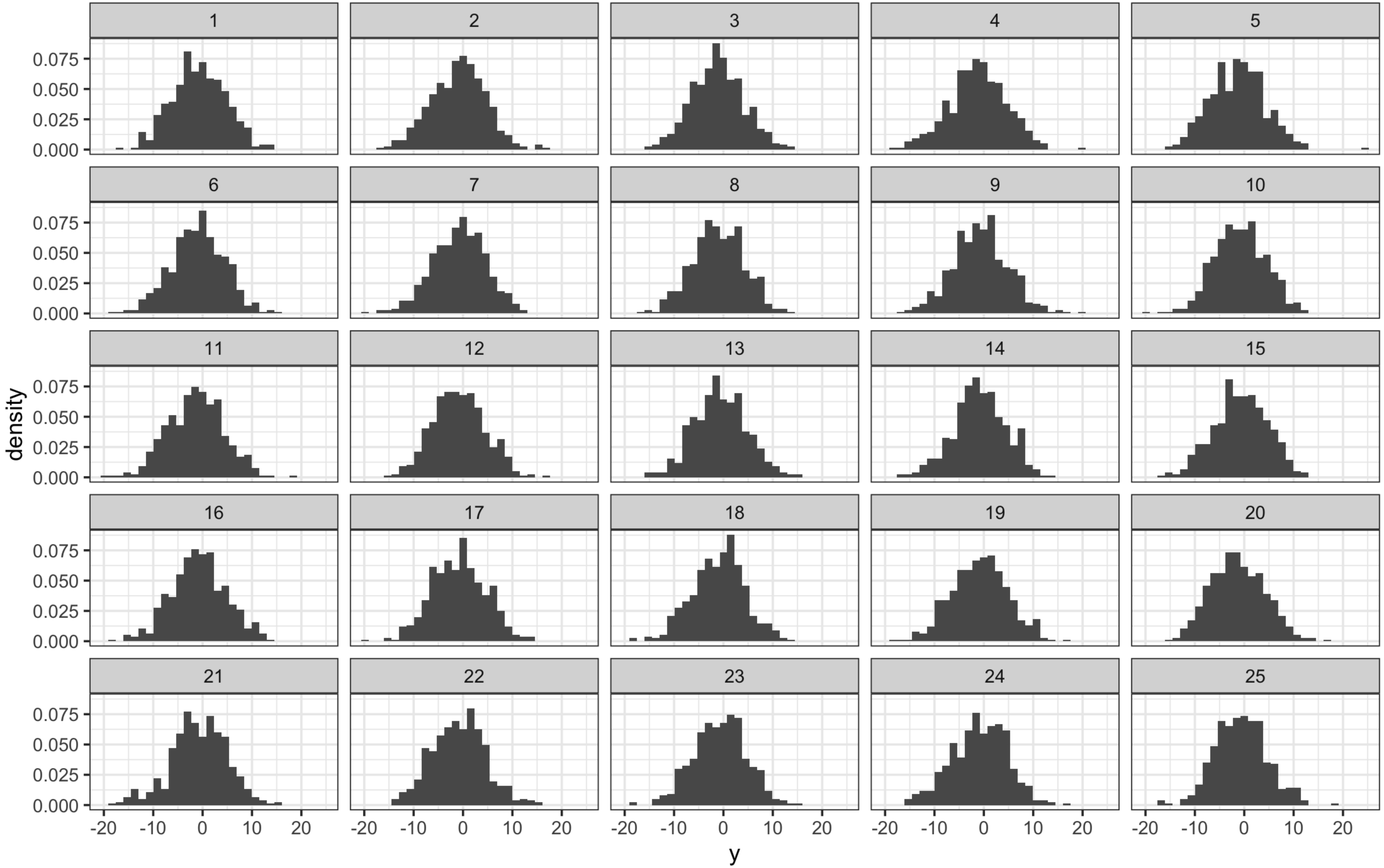
- What do the histograms of y look like?

  ↳ with $\beta_1 = -5$, the histograms look very not normally distributed

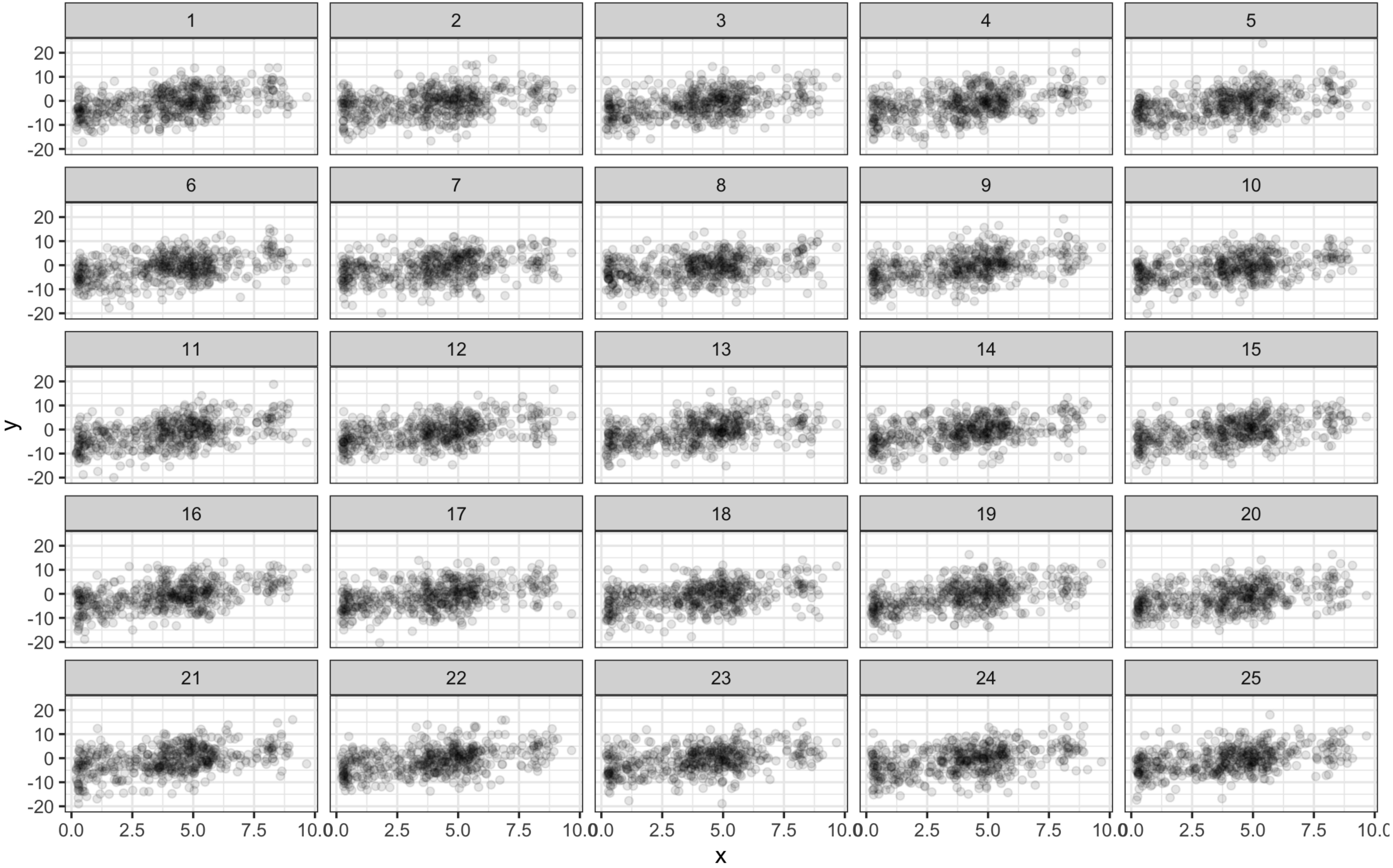
- What do the scatterplots of x vs y look like?

Example 3:

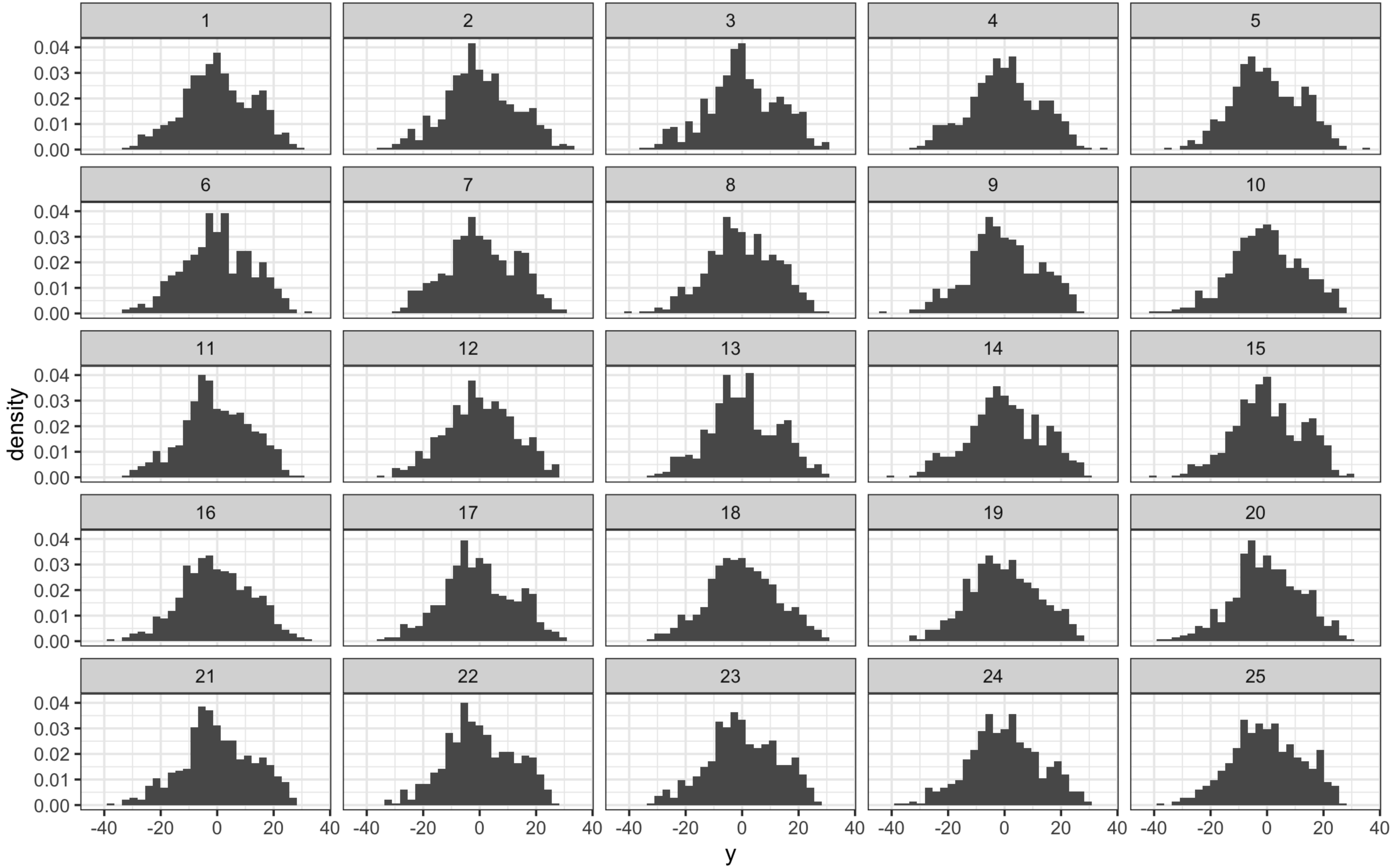
$$\{\beta_0 = -5, \beta_1 = 1, \sigma = 5\}$$ $\sigma^2 = 25$



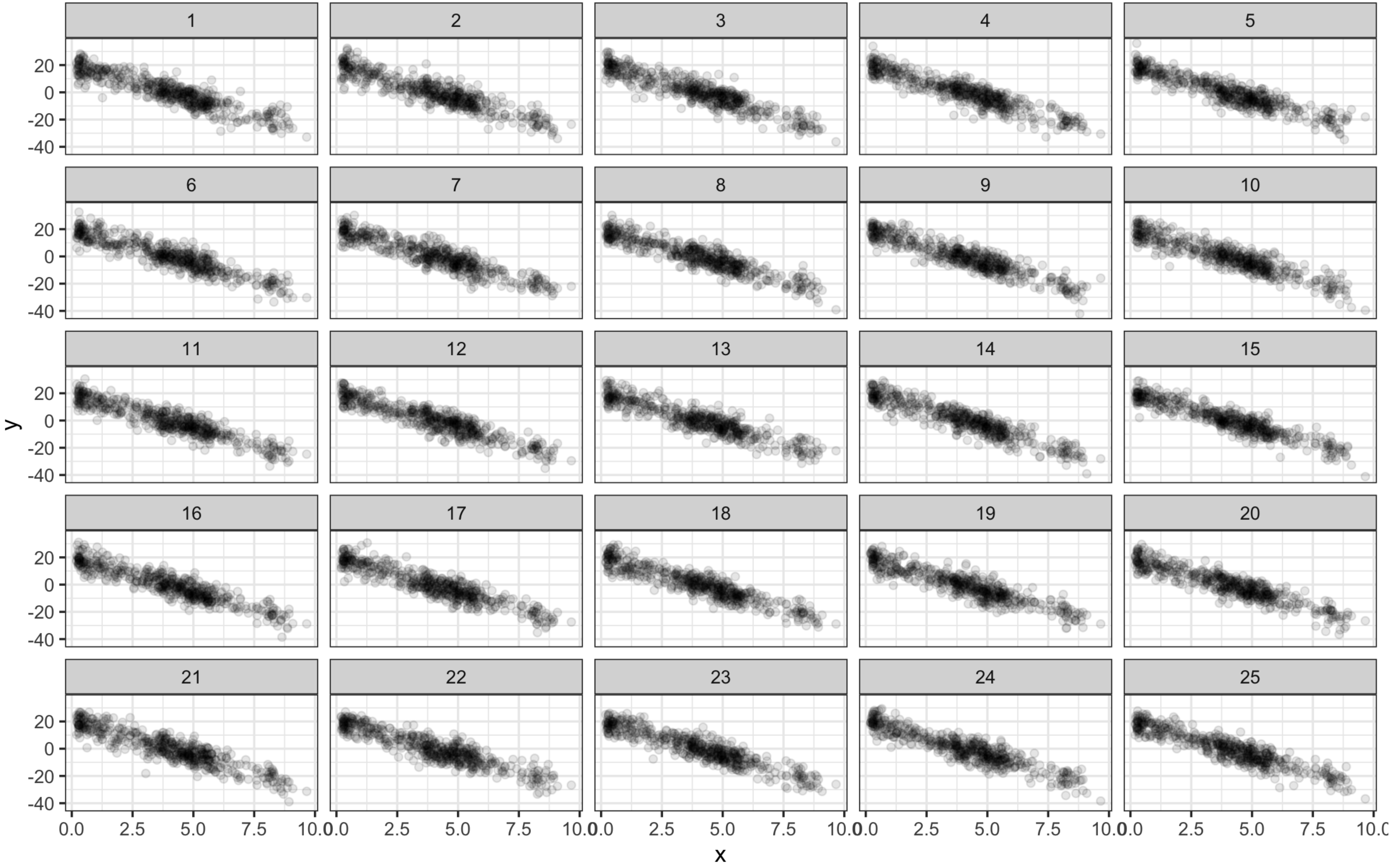
Example 3: $\{\beta_0 = -5, \beta_1 = 1, \sigma = 5\}$



Example 4: $\{\beta_0 = 20, \beta_1 = -5, \sigma = 5\}$



Example 4: $\{\beta_0 = 20, \beta_1 = -5, \sigma = 5\}$



What changed as we increased the value of σ?

- The two examples have the same trend, but we increased the value of σ… does that make a difference?

- What do the histograms of y look like?

- What do the scatterplots of x vs y look like?

Example 5:

$$\{\beta_0 = -5, \beta_1 = 1, \sigma = 20\}$$

$\sigma^2 = 400$

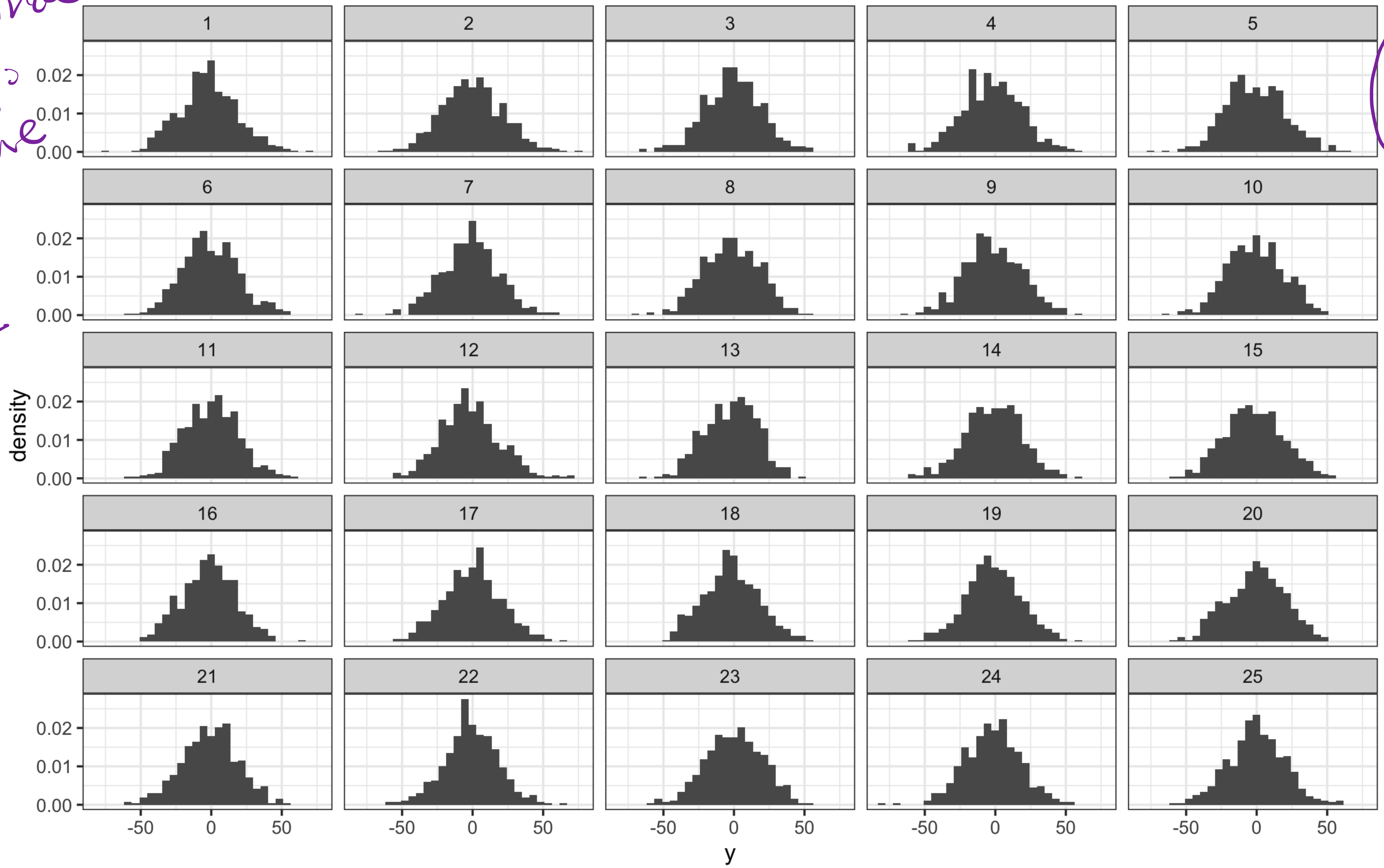
$\boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$

deterministic

stochastic (noise)

$\varepsilon_i \sim N(0, \sigma^2)$

the marginal dist of y_i doesn't have to be normal

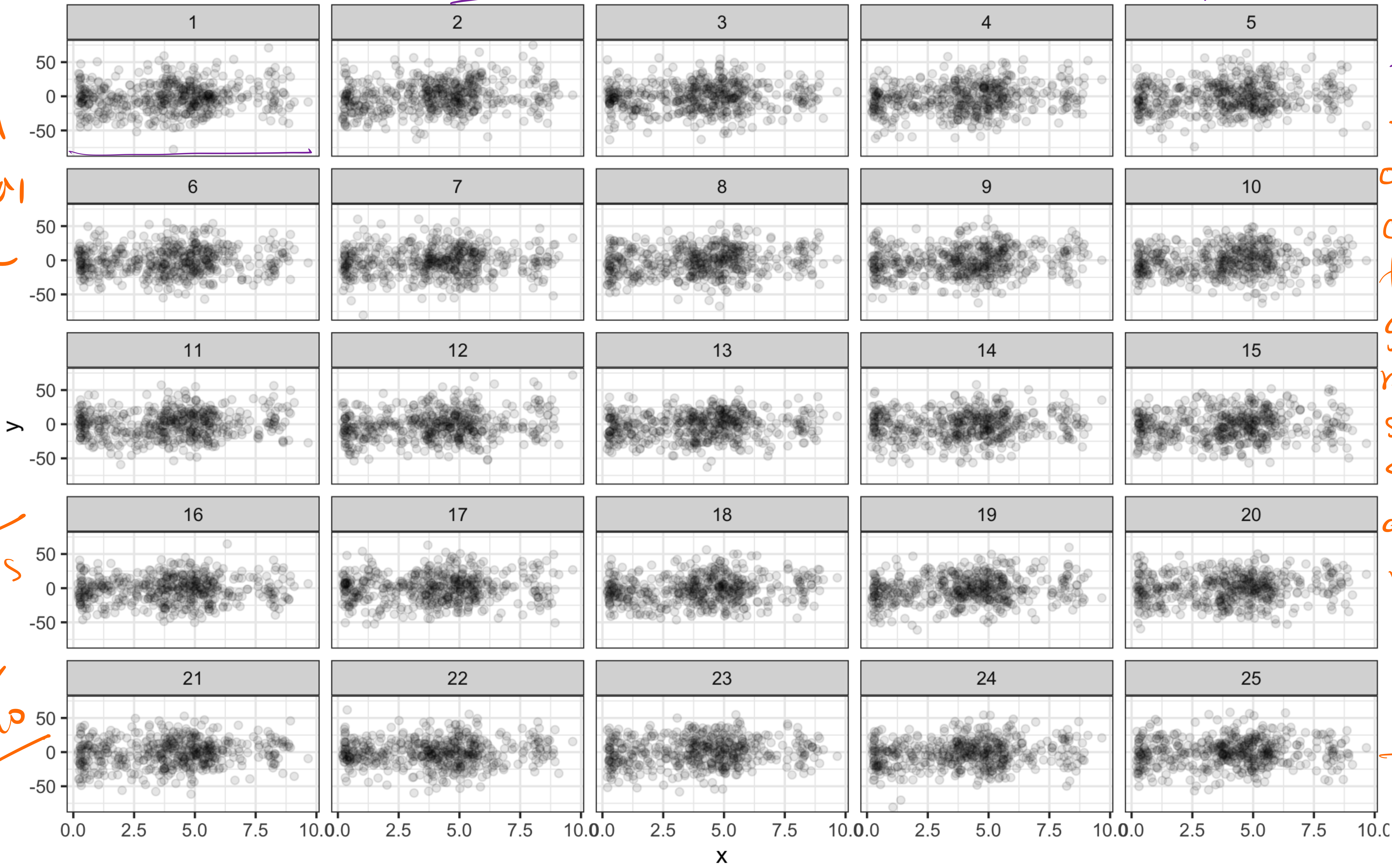


Example 5: $\{\beta_0 = -5, \boxed{\beta_1 = 1,} \sigma = 20\}$

the data below do follow a linear trend

if we only had the data, would we say that $\beta_1 > 0$?

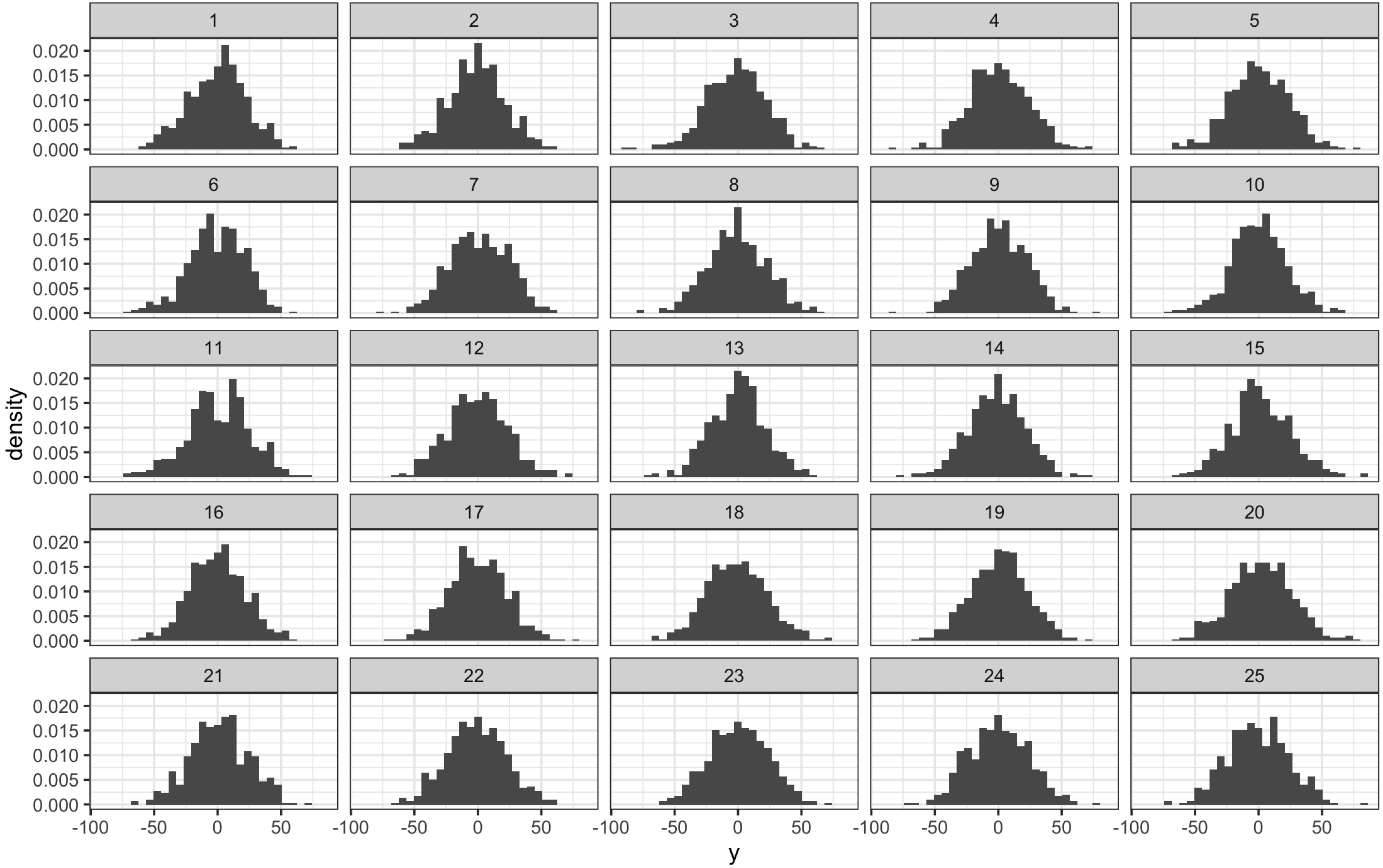
sometimes the answer is no



even if our data comes from a given model, sample size can affect what we can actually estimate.

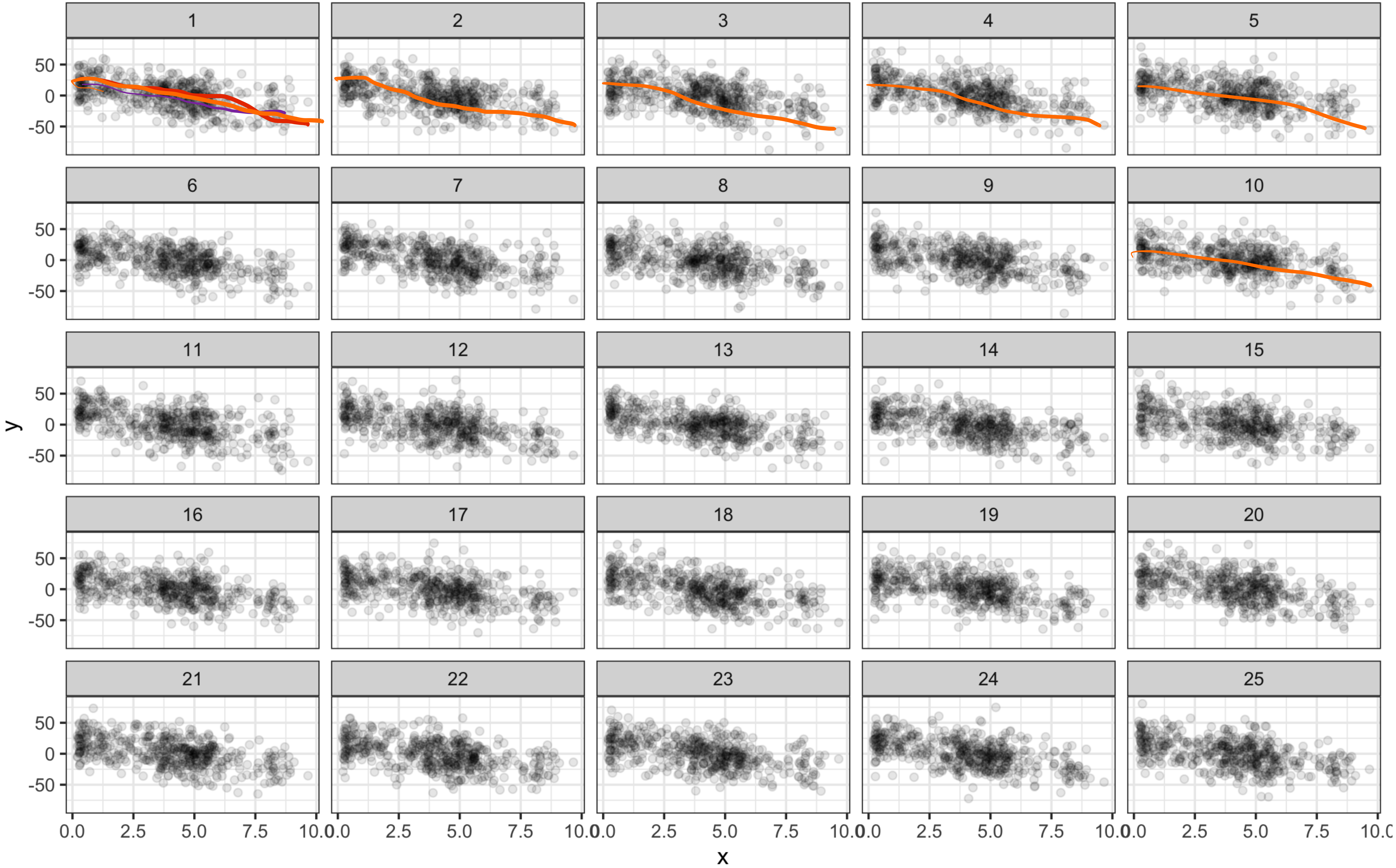
→ parameters

Example 6: $\{\beta_0 = 20, \beta_1 = -5, \sigma = 20\}$ $\sigma^2 = 400$



Example 6:

$$\{\beta_0 = 20, \beta_1 = -5, \sigma = 20\}$$



What do we learn by simulation?

- We learn what data looks like from a simple linear regression — the true meaning of using a statistical model as a 'data generating mechanism'.

- Even if the models have the same linear trend, $\beta_0 + \beta_1 x_i$, the value of σ can drastically change what our data looks like.

- How do we interpret the value of $\boxed{\beta_1}$ with varying values of $\boxed{\sigma}$?

Example models:

Example 1	$\beta_0 = -5, \beta_1 = 1$	$\sigma = 1$
Example 2	$\beta_0 = 20, \beta_1 = -5$	$\sigma = 1$
Example 3	$\beta_0 = -5, \beta_1 = 1$	$\sigma = 5$
Example 4	$\beta_0 = 20, \beta_1 = -5$	$\sigma = 5$
Example 5	$\beta_0 = -5, \beta_1 = 1$	$\sigma = 20$
Example 6	$\beta_0 = 20, \beta_1 = -5$	$\sigma = 20$

Extending the simple linear regression model

- Beyond simple linear regression, we can begin to incorporate more covariates to have:

conditional distribution

assuming a linear relationship

$$y_i \,|\, \boldsymbol{x}_i \sim N\left(\boxed{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij},}\ \sigma^2\right)$$

- Even further, we don't have to assume a linear relationship! Generalized additive models (GAMs) use flexible, nonlinear forms for modeling the relationship between x_j and y.

$$y_i = \beta_0 + f(x_i) + \epsilon$$

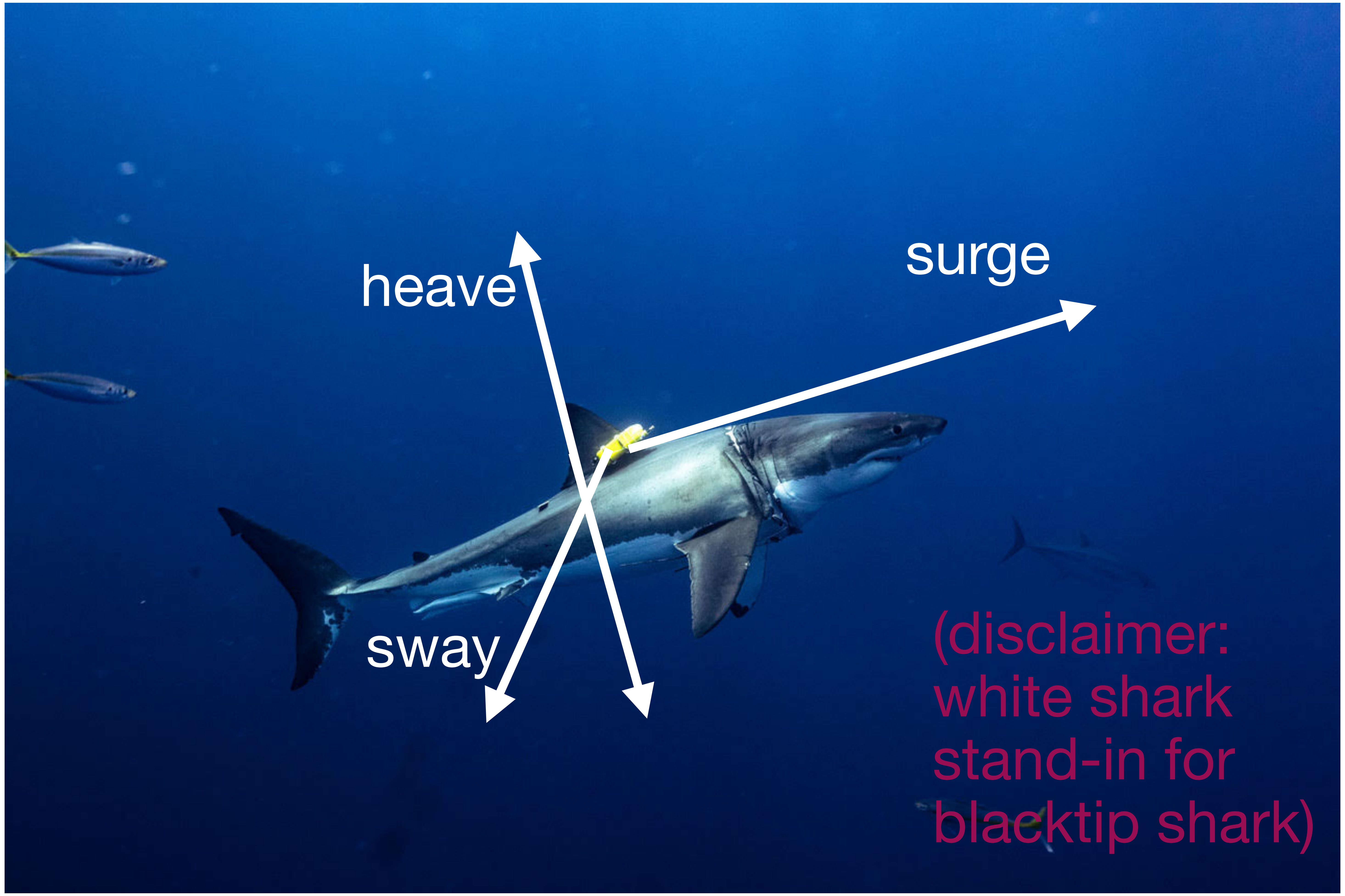
Example: Blacktip Reef Shark

Modeling overall dynamic body acceleration via linear regression

Measure: tri-axial acceleration *(heave, surge, sway)*
Compute: overall dynamic body acceleration (ODBA)

"Overall dynamic body acceleration is a single, integrated measure of body motion in all three spatial dimensions"

Paper: Making overall dynamic body acceleration work: on the theory of acceleration as a proxy for energy expenditure

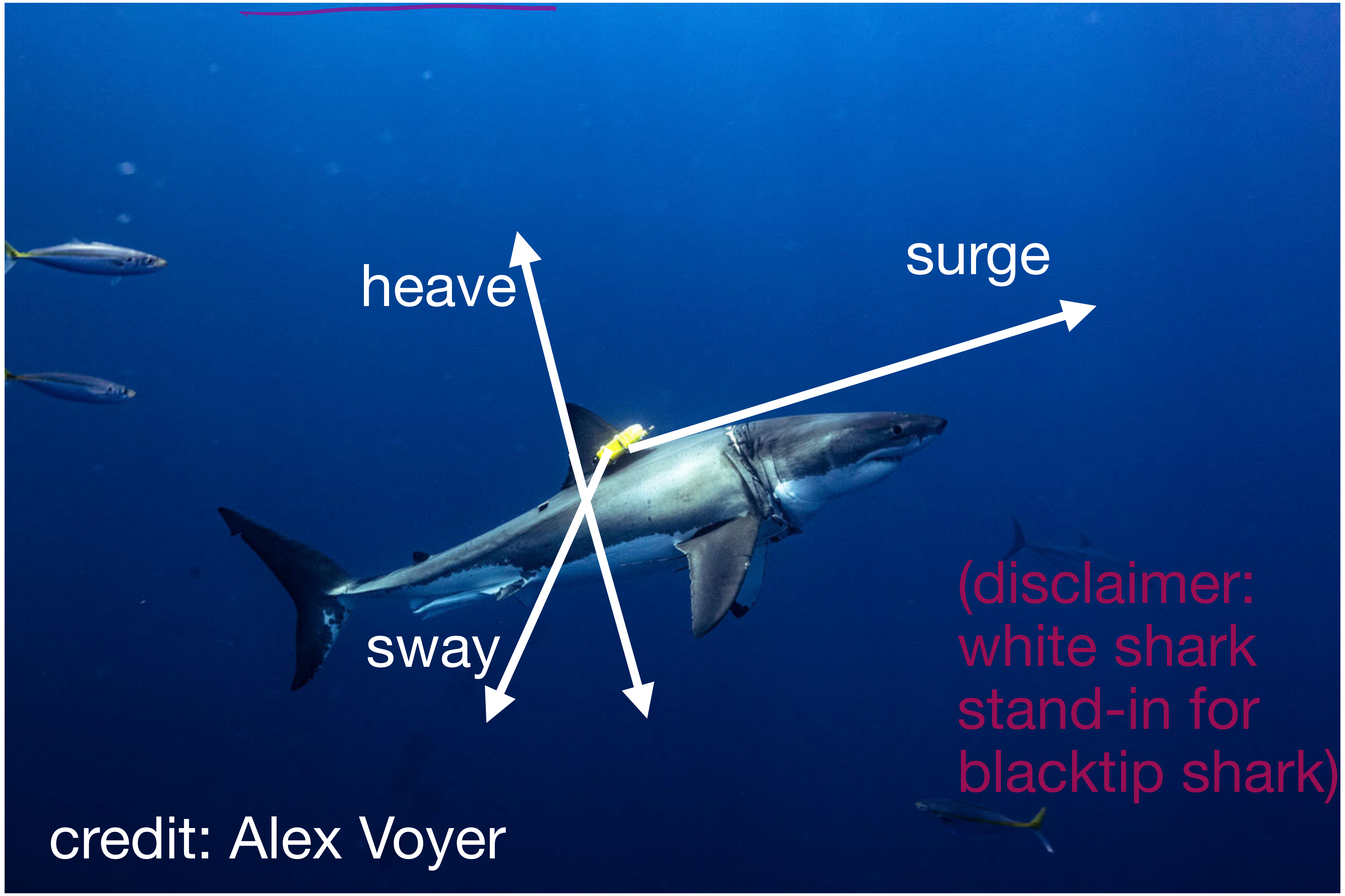


Accelerometer data is first decomposed into static and dynamic acceleration.

Then, we compute ODBA as the sum of the absolute values of the dynamic accelerations from all three axes.

gravity

$$ODBA = |heave_{dynamic}| + |surge_{dynamic}| + |sway_{dynamic}|$$



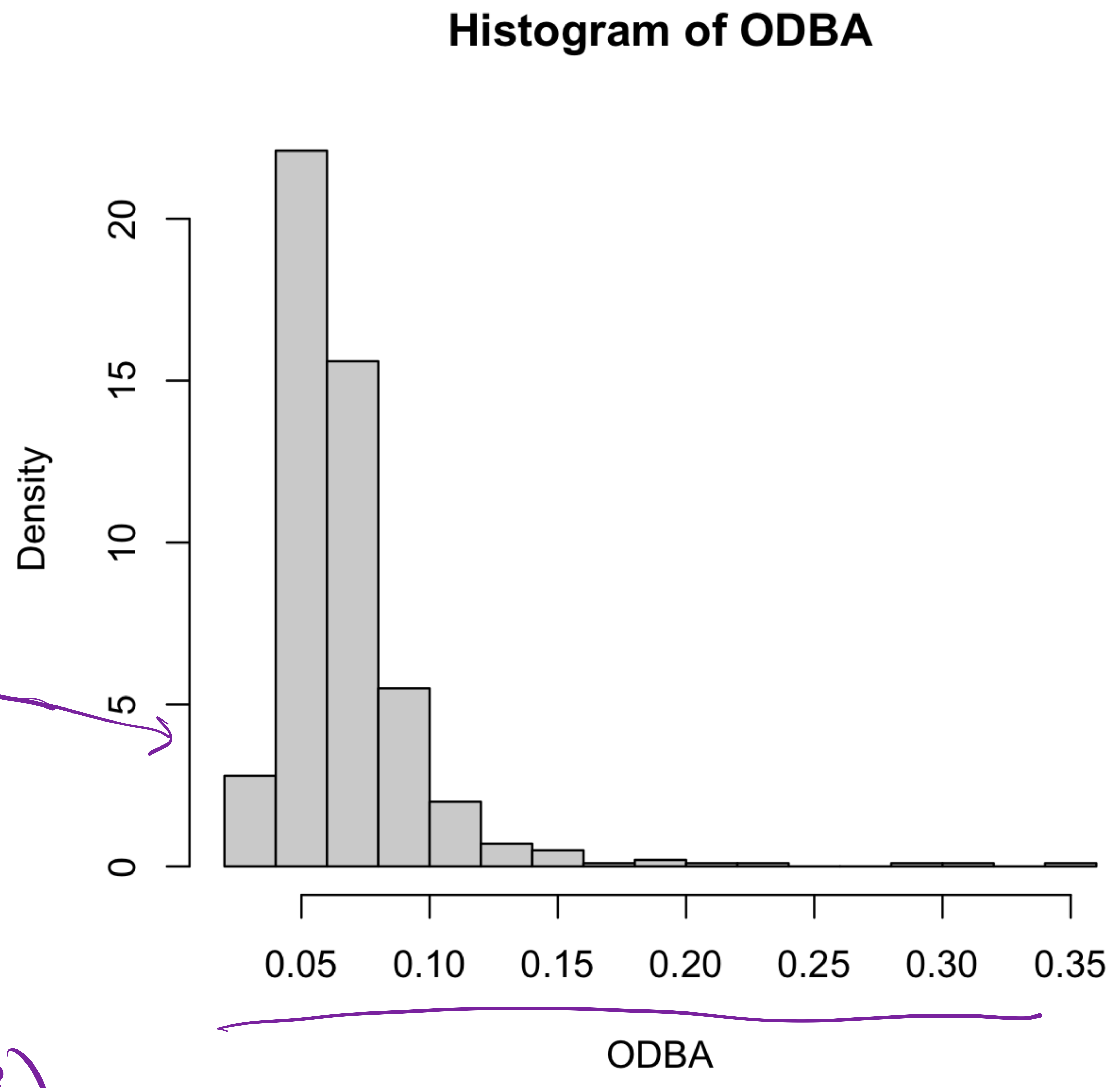
Random ODBA Samples

- Take 500 1-sec average measurements of ODBA
- Computed average depth (in m) for the shark
- What is the range of ODBA?
- Is this a problem?

$(0, \infty)$

$$ODBA = \beta_0 + \beta_1 \cdot depth + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$



$$ODBA \mid depth \sim N(\beta_0 + \beta_1 depth, \sigma^2)$$

Back to Simple Linear Regression

$\in \mathbb{R}$ but ODBA on $(0,\infty)$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad for \quad i = 1,2,\ldots,I$$

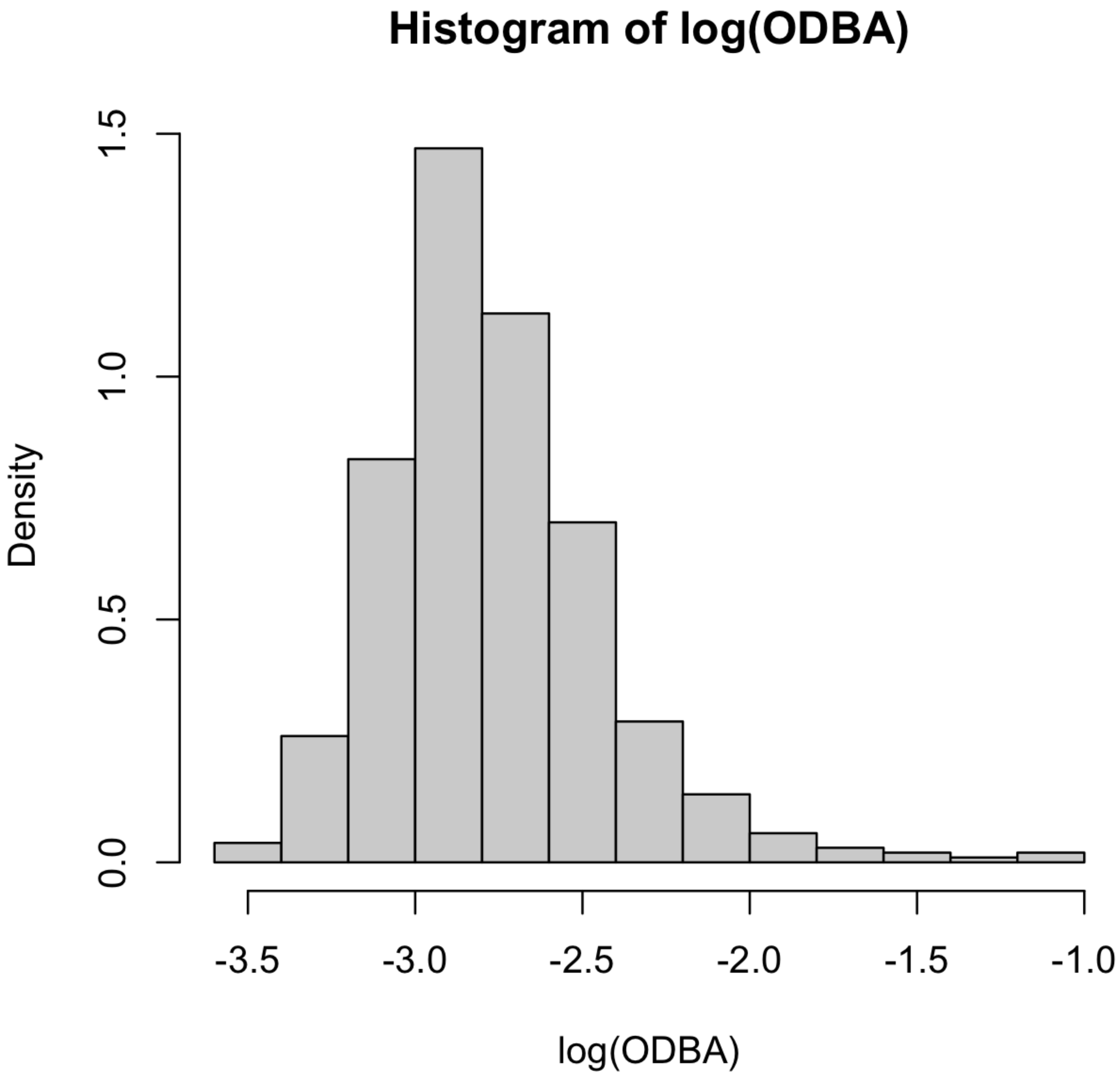
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Suppose $y_i \in (0,\infty)$, then depending on the domain of x_i, certain values of β_0 and β_1 as well as σ may lead to predicted values of $y_i < 0$.

We *a priori* assume that $\beta_0, \beta_1 \in \mathbb{R}$, which may not be the case here. To account for this, we could construct a model for $log(ODBA)$. $\in \mathbb{R}$

Random log(ODBA) Samples

- Take 500 1-sec average measurements of ODBA and compute log(ODBA)
- Computed average depth (in m) for the shark
- What is the range of log(ODBA)?
- Slide ~~11~~ shows the distribution of depth that we'll use as a covariate

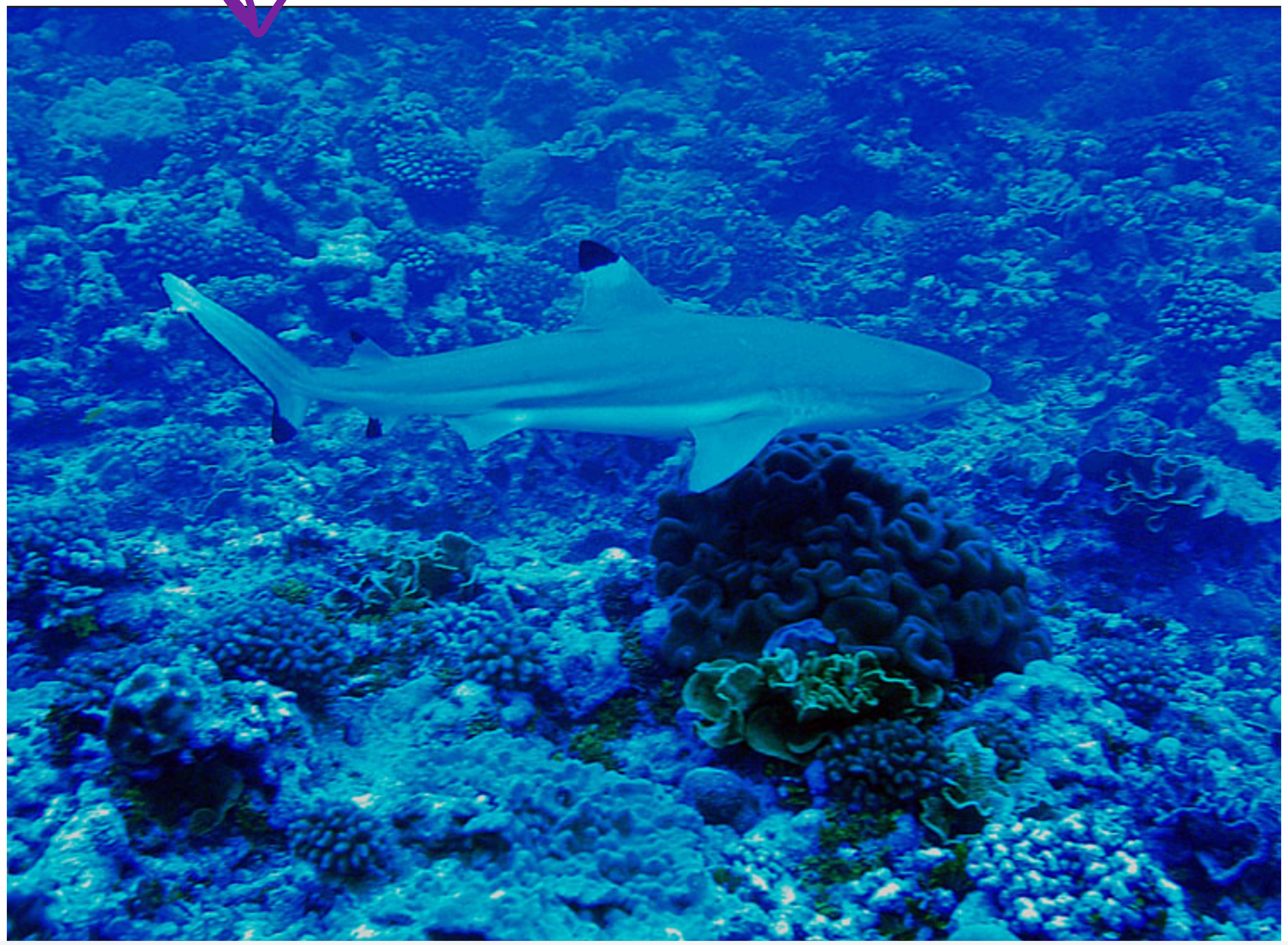


log(ODBA) vs Depth

Is there a linear relationship between log(ODBA) and depth?

$$log(ODBA)_i = \beta_0 + \beta_1 depth_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$



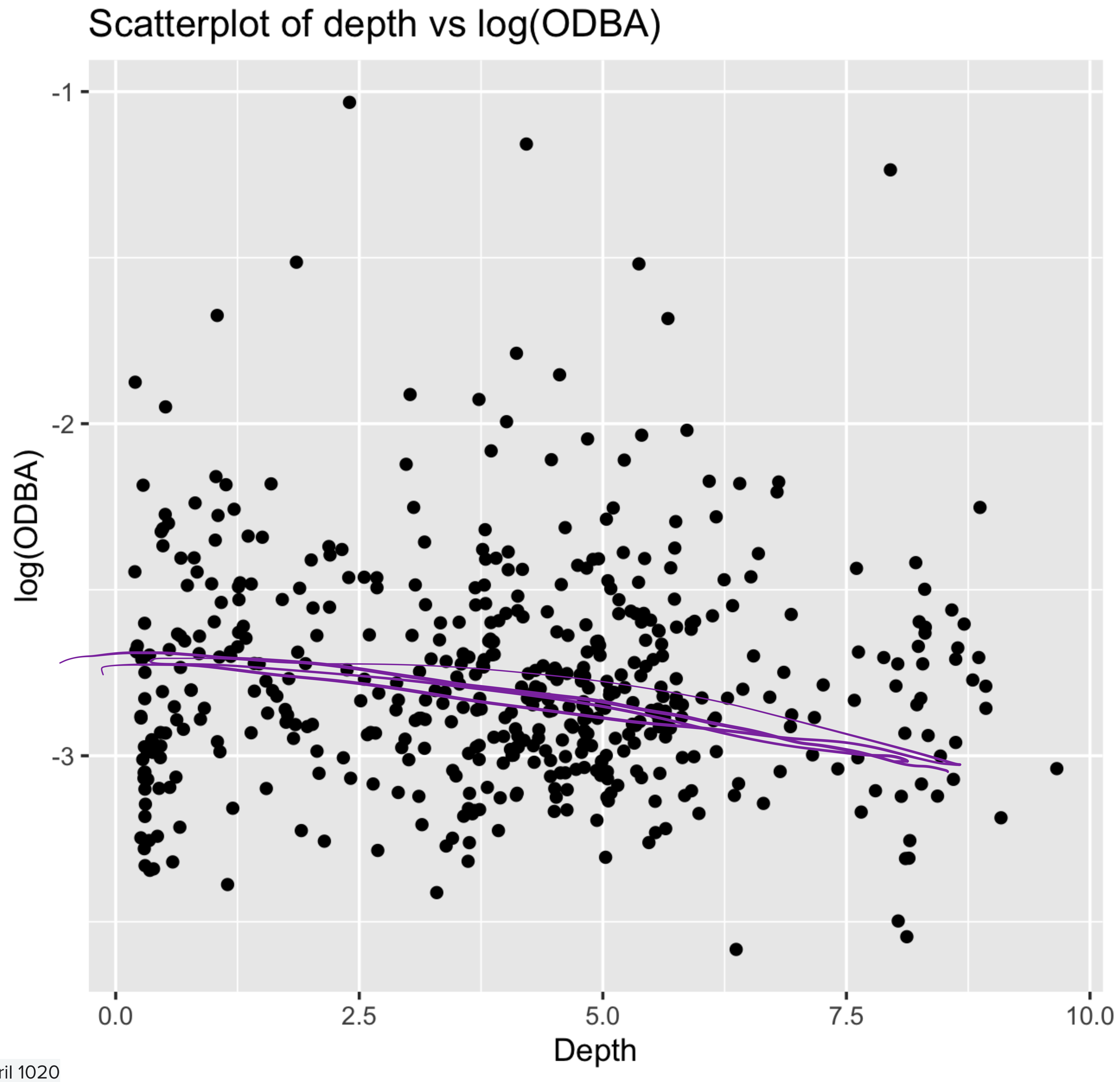


Photo Credit Laura M. Beauregard, USFWS - Blacktip reef sharak (Carcharhinus melanopterus) on the north side of the Palmyra Atoll. Date: April 1020

Generalized linear models (GLMs)

What if we don’t have continuous data?

- In SLR, we make a few assumptions about the process of interest, specifically that the observation process is continuous and defined on the real line but this allows us to understand the (linear) relationship between x and y.

- In practice, there’s lots of x_i processes that result in data that are different and we may want to include covariates to understand how they’re connected:

  - how do counts vary (e.g. how do sloth populations vary in Costa Rica across years?)
  - what affects the probability (e.g. under what conditions are we more likely to spot a whale?)

- We learned about other distributional forms and we can extend those!

General GLMs

- In addition to the deterministic (systematic), stochastic (random) components of a linear model, a generalized linear model will typically have a third component: a link function.

- The purpose of the link function is to map parameters on a constrained space onto the real line.

- ~~What was the link function for the Bernoulli distribution?~~

Data generated by a GLM

- In a regression context, while much of the focus is put on the deterministic/systematic component, we **should care more about the distributions**.

- What distributions are implied under different values rather than just what is the expected value under different values of the covariates

Extending the Bernoulli distribution

parameter

$$y_i \sim Bern(p)$$

- We want to let the parameter be a function of the presence of some covariate.

- We know that $\boxed{p}$ is bounded between 0 and 1, so does the following work?

probability varies depending on the value of s

$$p = \beta_0 + \beta_1 s$$

for some values of β_0 and β_1, we can end up with values of p that fall outside of (0,1) logistic regression

Extending the Bernoulli distribution

we use a link function to be able to do linear regression

- Instead, we use a logit link for p so the full data model is given as:

$$y_i \sim Bern(p_i)$$

logit link maps $(0,1) \to \mathbb{R}$

$$p_i = \frac{\exp(\beta_0 + \beta_1 s_i)}{1 + \exp(\beta_0 + \beta_1 s_i)}$$

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 s_i$$

— $\beta_0, \beta_1 \in \mathbb{R}$

fixed

- So, what is the conditional distribution of y_i given s_i here? What is the expected value and variance?

$y_i | s_i \sim Bern(p_i)$

Example: horseshoe crab

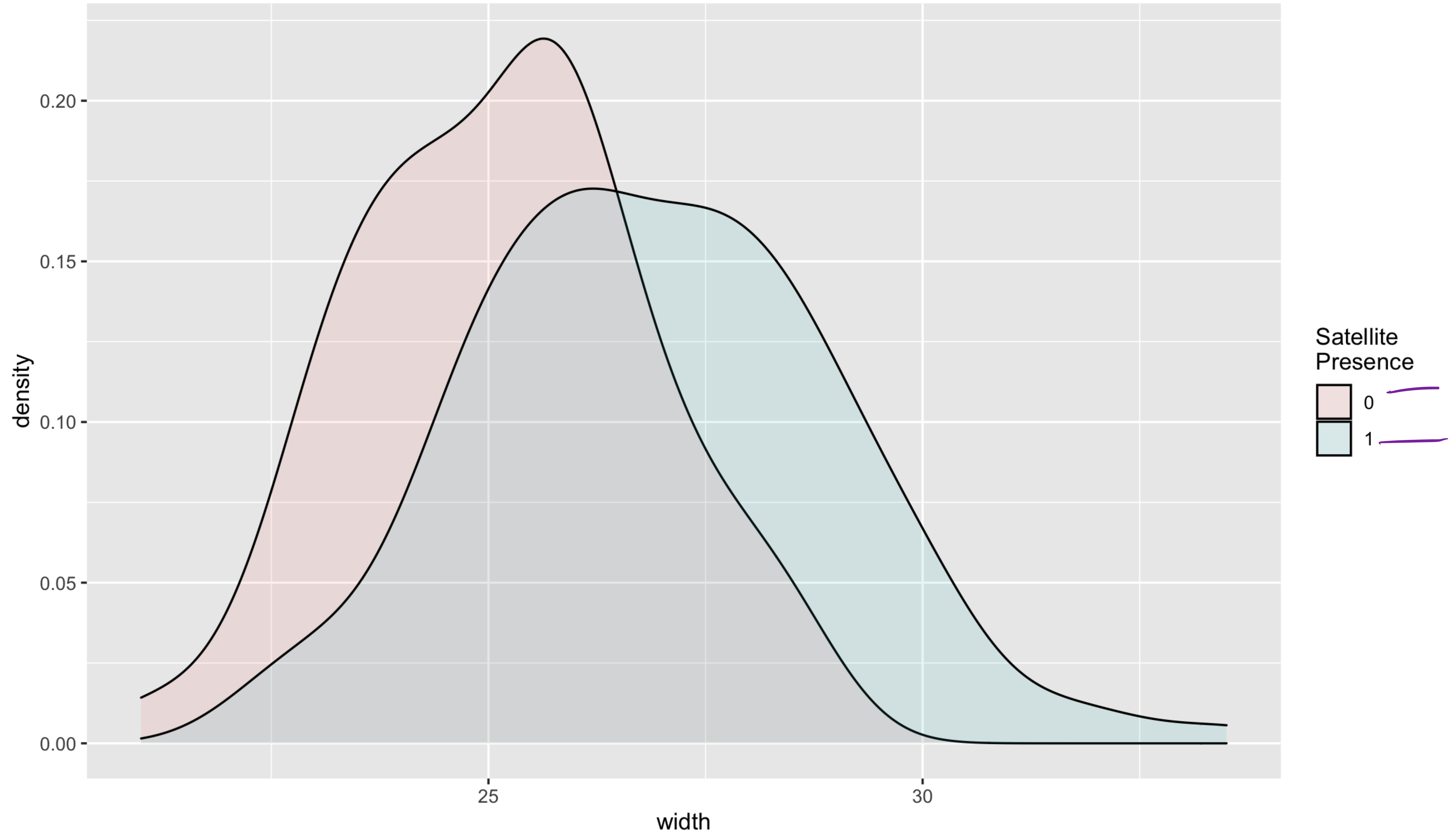
Looking at a study of 173 nesting horseshoe crabs

Each female horseshoe crab had a male crab attached to her in her nest. This study investigated factors that affected whether other males were (called satellites) residing nearby her.

- y = whether a female crab has a satellite (1 = yes, 0 = no)
- width = female horseshoe crab's shell width in cm

Brockmann, H. J. 1996. Satellite male groups in horseshoe crabs. *Limulus polyphemus*. Ethology 102: 1-21

Satellite presence vs shell width



Constructing a model for the horseshoe crabs

- Our random variable of interest is the presence of the satellites, y_i, and y_i can either be 0 or 1, so we need to assign a distribution that places probabilities on either of these two values

- We can assign y_i a Bernoulli distribution with parameter p_i which tells us the probability of a '1' and $1 - p_i$ is the probability of a '0'

$$y_i \sim Bern(p_i)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 w_i$$

- What is the expected value and variance here for a given w_i?

Poisson distribution in a GLM framework

- Random component: $y_i \sim Pois(\lambda)$, $\lambda > 0$

- Link function: $log(\lambda)$ $\in \mathbb{R}$

- Systematic component: $log(\lambda) = \beta_0 + \cdots$

Modeling count data

- Count data (e.g. data that takes on values in 0, 1, 2, ….) can be modelled via the generalized linear modeling framework

∞

- We can take our random variable y_i to reflect the counts and use a Poisson distribution but another common distribution for counts is the negative binomial distribution.

$$y_i \sim Pois(\lambda) \qquad y_i \sim NB(\mu, \phi)$$

Poisson vs Negative Binomial Distribution

Poisson

- $y_i \sim Pois(\lambda)$
- $E(y_i) = \lambda$; $Var(y_i) = \lambda$
- $\lambda > 0$

if $\lambda = \mu$

Negative Binomial

- $y_i \sim NB(\mu, \phi)$
- $E(y_i) = \mu$; $Var(y_i) = \mu + \mu^2/\phi$
- $\mu > 0$; $\phi > 0$

As $\phi \to \infty$, the negative binomial distribution converges to the Poisson distribution.

Poisson vs Negative Binomial Distribution

- If we take $\mu = \lambda$, we can see that the negative binomial distribution allows for a larger amount of variation than we would expect from a Poisson distribution.

$$E(y_i) = \lambda \qquad\qquad Var(y_i) = \lambda + \lambda^2/\phi$$

- For instance, if $\lambda = 5$, the Poisson distribution would give us

$$E(y_i) = 5 \qquad\qquad Var(y_i) = 5$$

and the negative binomial distribution would give us

$$E(y_i) = 5 \qquad\qquad Var(y_i) = 5 + 25/\phi$$

GLM Framework

- **Random component:** $y_i \sim Pois(\lambda_i)$; $y_i \sim NB(\mu_i, \phi)$

  $\lambda_i > 0$ $\mu_i > 0$

- **Link function:** $log(\lambda_i)$; $log(\mu_i)$

  $\in \mathbb{R}$ $\in \mathbb{R}$

- **Systematic component:** $log(\lambda_i) = \beta_0 + \cdots$; $log(\mu_i) = \beta_0 + \cdots$

Note that we often allow covariates to inform the mean structure, but in some cases the variance terms can also be a function of covariates.

(in a linear regression framework, we often include covariates in the mean component)

e.g. $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

constant

Again, we say nothing about the marginal dist of y_i

the $y_i | x_i \sim Pois(\lambda_i)$

$log(\lambda_i) = \beta_0 + \beta_1 x_i$

$\longrightarrow y_i | x_i \sim Pois(e^{\beta_0 + \beta_1 x_i})$

Modeling Counts

- Parameters of the Poisson distribution: λ
  - $\lambda > 0$, reflects the mean and variance of the process

- Parameters of the negative binomial distribution: μ, ϕ
  - $\mu > 0, \phi > 0$, where μ reflects the mean of the process and ϕ reflects the difference in variance between the Poisson distribution and the negative binomial distribution (larger ϕ leads to properties similar to a Poisson)

In the code….

- Let's see if you can extend the Poisson and negative binomial distributions to include covariates in the mean structure.

- Simulate covariates and produce replicate data sets for each extension.

Overall

- The main goal of today has been to understand:

  - what a probability distribution/mass function is
  - understand the role they play in building statistical models
  - learning to understand what a statistical model actually does by generating data from it

- Data generated from a statistical model may not always look ‘neat’ or how we expect it to look like. That’s part of building your intuition.

- At the end of the day, working in interdisciplinary settings requires we communicate across disciplines and simulation/producing ‘fake data’ has been one of the easiest ways to ‘speak’ the same language.