Stat310

Sampling distributions

Hadley Wickham

1. Stats in practice & test
2. Chebyshev clarification
3. LLN completion
4. CLT & approximations
5. Sampling distributions

Assessment

Stats in practice

Graded and returned. If you got <5 you may resubmit without penalty – make sure to write one page and answer all the three questions!

Will have rubrics and samples up before next one.

Test

Available Mar 22. Due Mar 29.

Same format as last time, and aiming for about the same difficulty.

120 minute take home test. 4 questions.

Approximately half applied (working with real problems) and half theoretical (working with mathematical symbols).

Material

Covers everything up to Mar 20: transformations, bivariate random variables, sequences and sampling distributions. See website (soon) for exactly what you should know.

Honour code

No collaboration.

No communication about the questions or your answers.

Only outside resources allowed are: a one-page double-sided note sheet and wolfram alpha.

Pledged and signed.

Expectations

Points will be awarded for fully converting a word problem into a mathematical problem.

You should be able to recall any fact on the basic math sheet.

I will supply random mathematical facts and tables of probabilities (if needed). You may use wolframalpha, but it will not be necessary.

Chebyshev

$$P(|X - \mu| > K\sigma) \leq \frac{1}{K^2}$$

A bit confusing

$$P(|X - \mathrm{E}(X)| > K\sqrt{\mathrm{Var}(X)}) \leq \frac{1}{K^2}$$

More explicit

$$P(|X_n - \mathrm{E}(X_n)| > K\sqrt{\mathrm{Var}(X_n)}) \leq \frac{1}{K^2}$$

$$P(|X_n - \mu| > K\sqrt{\frac{\sigma^2}{n}}) \leq \frac{1}{K^2}$$

LLN

Example

You have a loaded die that has probability p of rolling a six, and (1 - p)/5 of rolling any other number.

What sequence of random experiments could you use to find p?

CLT

CLT

sample statistic

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$\lim_{n\to\infty} Z_n = Z$$

$$Z \sim N(0,1)$$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

$$\bar{X} \xrightarrow{d} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Which one of these does not make sense?

New notation

If $x_n \to 0$, and n is big, we can say $x_n \sim 0$.

If $X_n \to Z$, $Z \sim N(0, 1)$, and n is big,
we can say $X_n \sim\sim N(0,1)$.

Read as **approximately distributed**.

Conditions

$$X_i \overset{iid}{\sim} [\mu, \sigma^2]$$

Central limit **theories**

Actually, lots of different central limit theories. Basically vary in the conditions that they put on the X's.

(More conditions = easier to prove. Fewer conditions = applies to more situations)

We'll prove two next week.

Your turn

What's the difference between the LLN and the CLT?

Sampling distributions

Random experiment

"A random experiment is an experiment, trial, or observation that can be **repeated numerous times under the same conditions**... It must in no way be affected by any previous outcome and cannot be predicted with certainty." (http://cnx.org/content/m13470/latest/)

Where we are

Univariate random variables:
an experiment with **one** output

Bivariate random variables:
an experiment with **two** outputs.

Sequences of random variables:
An experiment performed repeatedly.
Repeatable = i.i.d

A **sampling distribution**: distribution of a summary statistic of a repeated experiment

Definitions

Sample = results of **n** random experiments.

Random sample = result of a random experiment **repeated** n times. Therefore, they're iid.

Both are sequences of random variables.

Statistic = A function of random variables with no unknown parameters.

Example statistics

Sample mean

Proportion of successes

Minimum, median, maximum

Standard deviation, variance

Example

Spin a bottle and record the angle in degrees in which it points. Repeat.

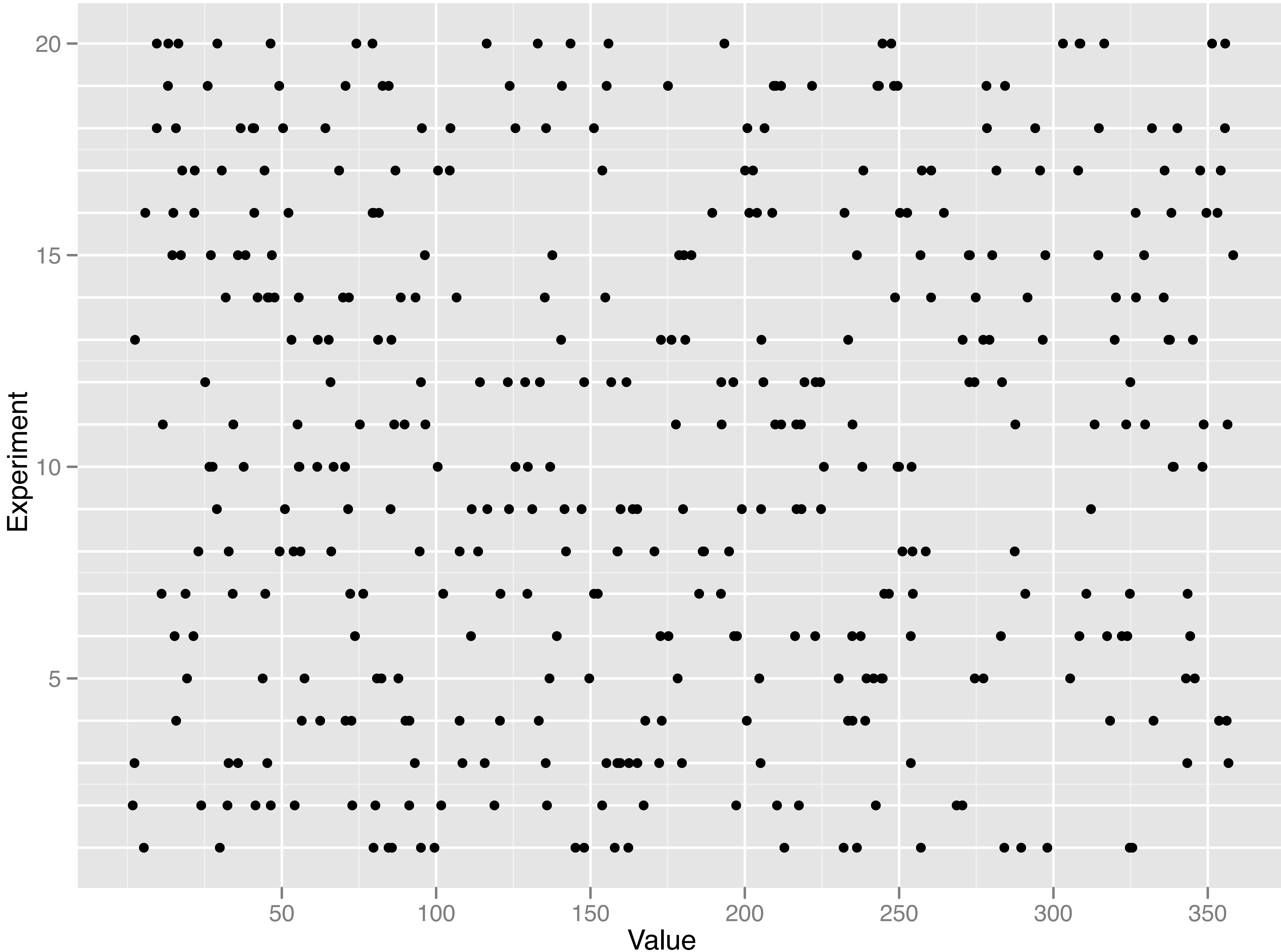
How would you write this mathematically?

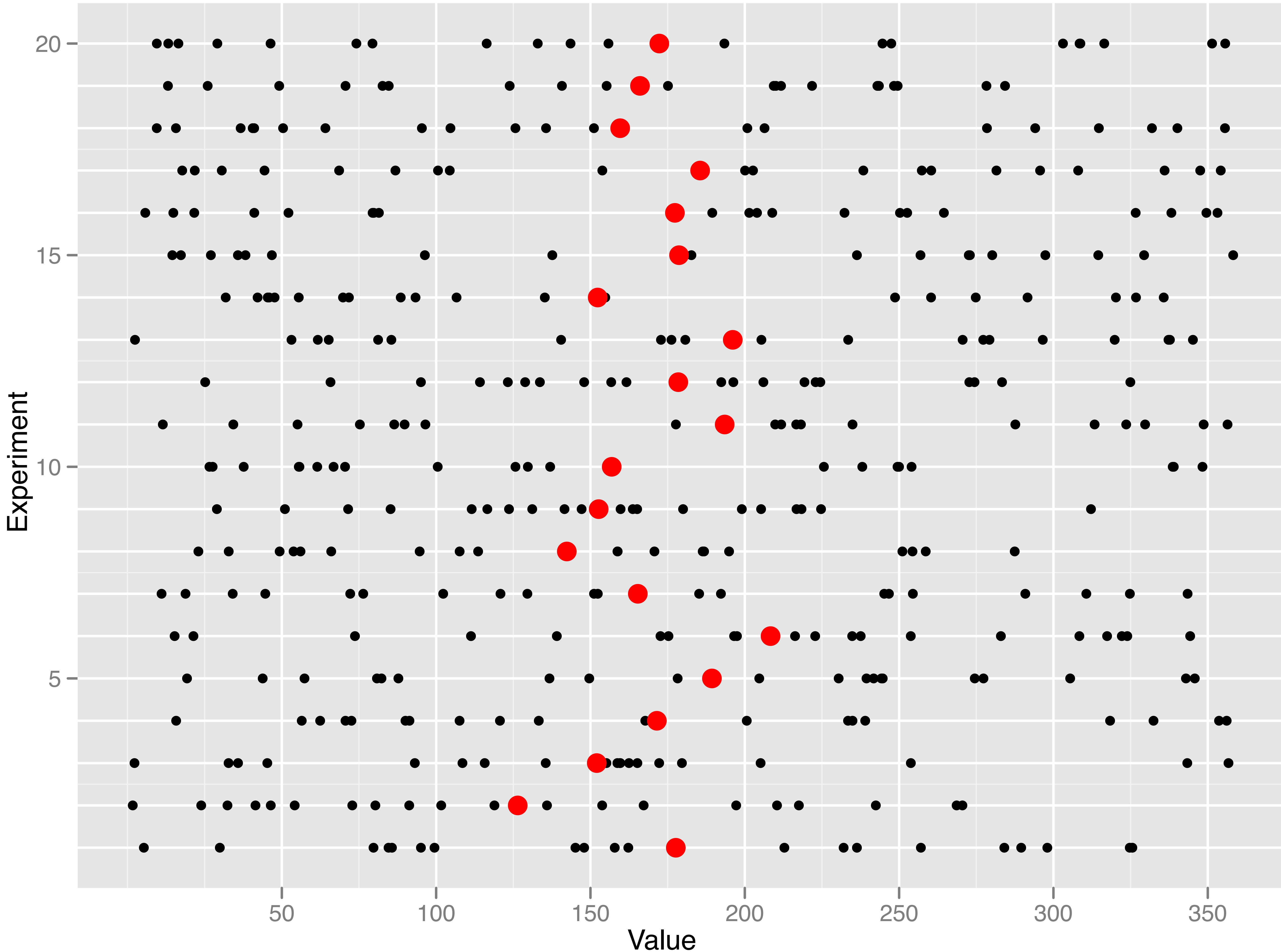
First time

$x_1 = 205$, $x_2 = 256$, $x_3 = 86$, $x_4 = 119$,
$x_5 = 16$, $x_6 = 278$, $x_7 = 55$, $x_8 = 16$,
$x_9 = 295$, $x_{10} = 341$, $x_{11} = 299$, $x_{12} = 270$,
$x_{13} = 118$, $x_{14} = 360$, $x_{15} = 97$, $x_{16} = 282$,
$x_{17} = 42$, $x_{18} = 283$, $x_{19} = 259$, $x_{20} = 326$

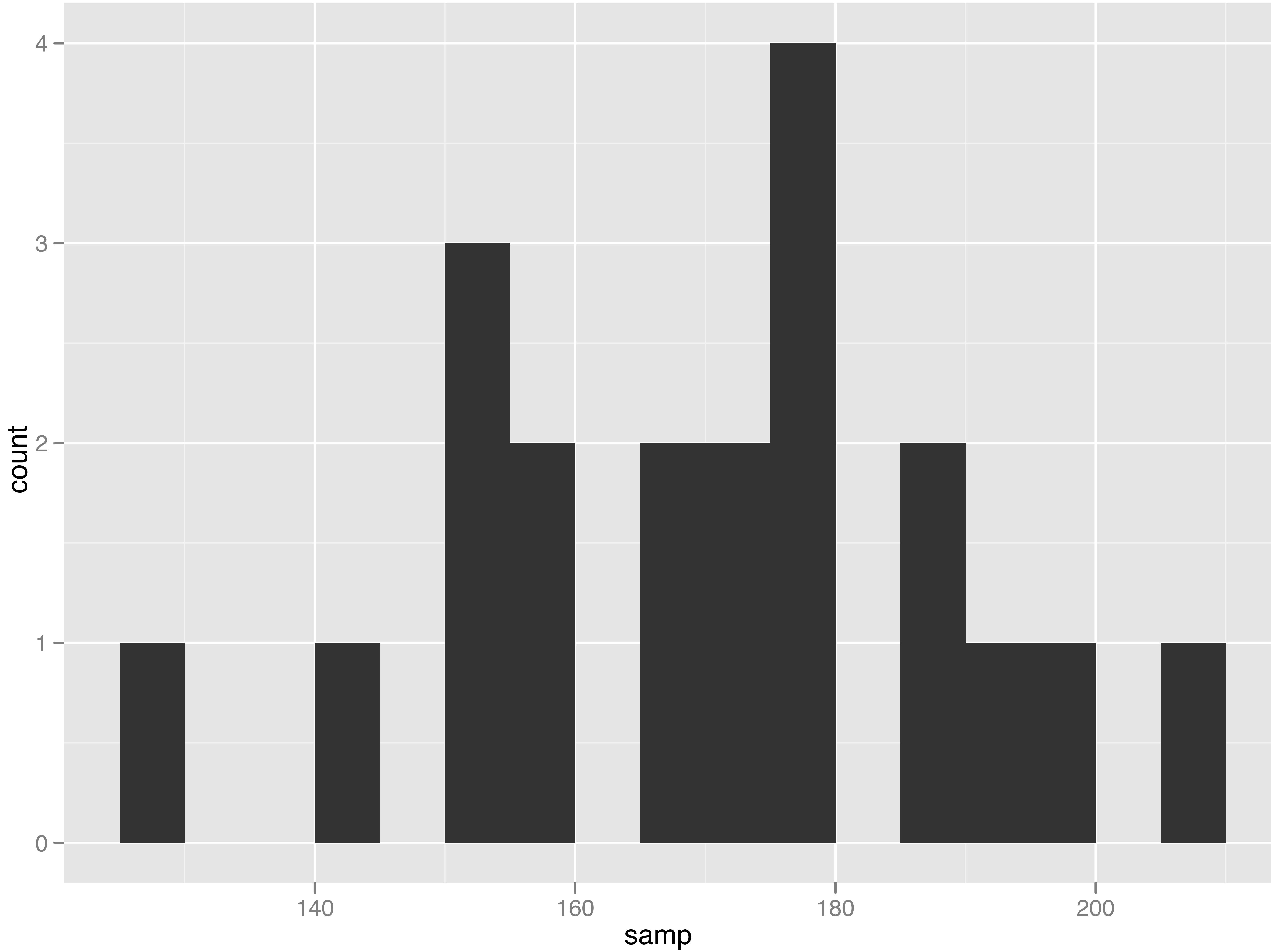
Second time

$x_1 = 184$, $x_2 = 344$, $x_3 = 118$, $x_4 = 226$,
$x_5 = 208$, $x_6 = 106$, $x_7 = 332$, $x_8 = 310$,
$x_9 = 339$, $x_{10} = 95$, $x_{11} = 7$, $x_{12} = 274$,
$x_{13} = 120$, $x_{14} = 346$, $x_{15} = 211$, $x_{16} = 166$,
$x_{17} = 84$, $x_{18} = 102$, $x_{19} = 32$, $x_{20} = 128$

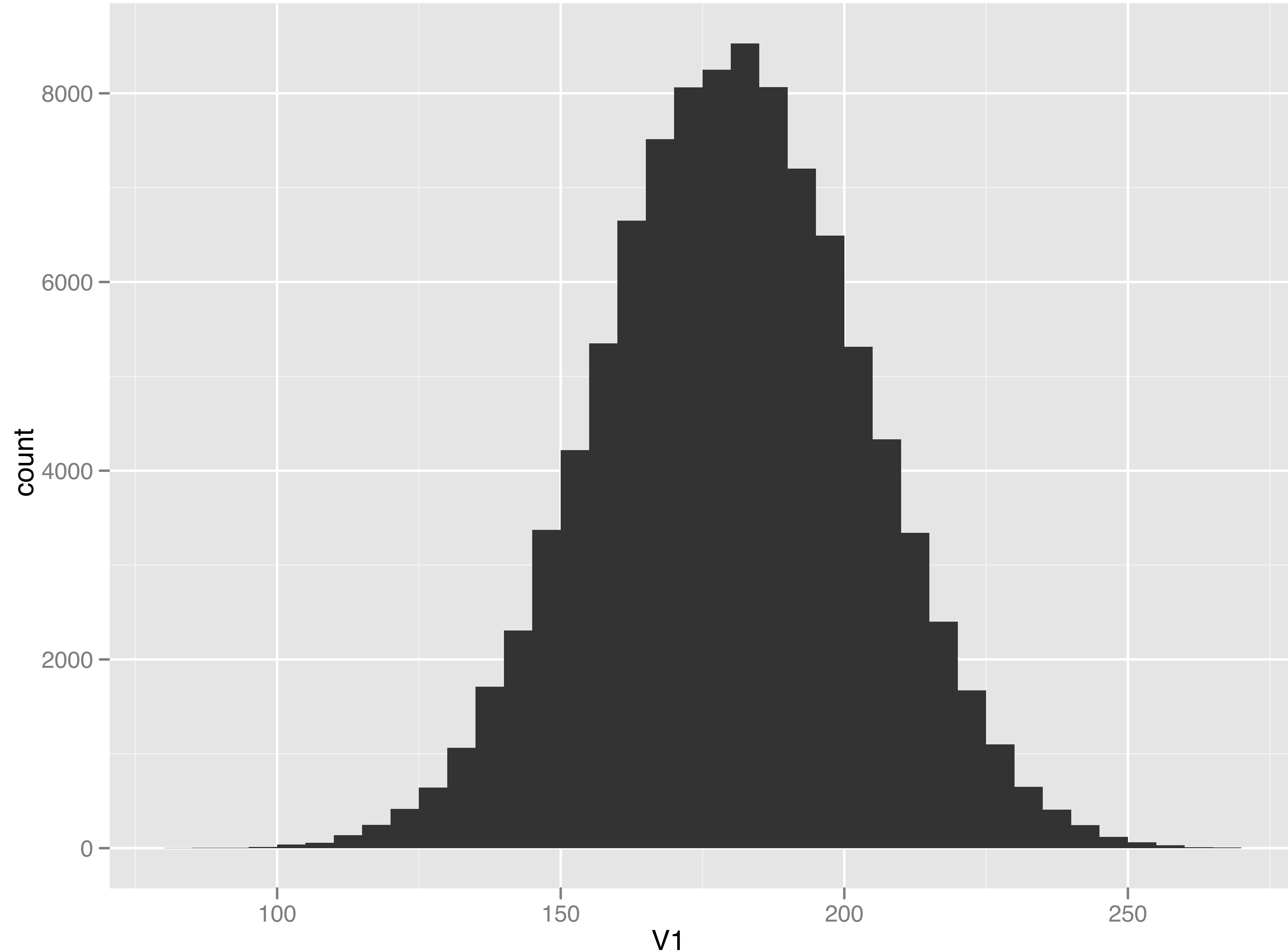

Experiment
Value
20
15
10
5
50
100
150
200
250
300
350

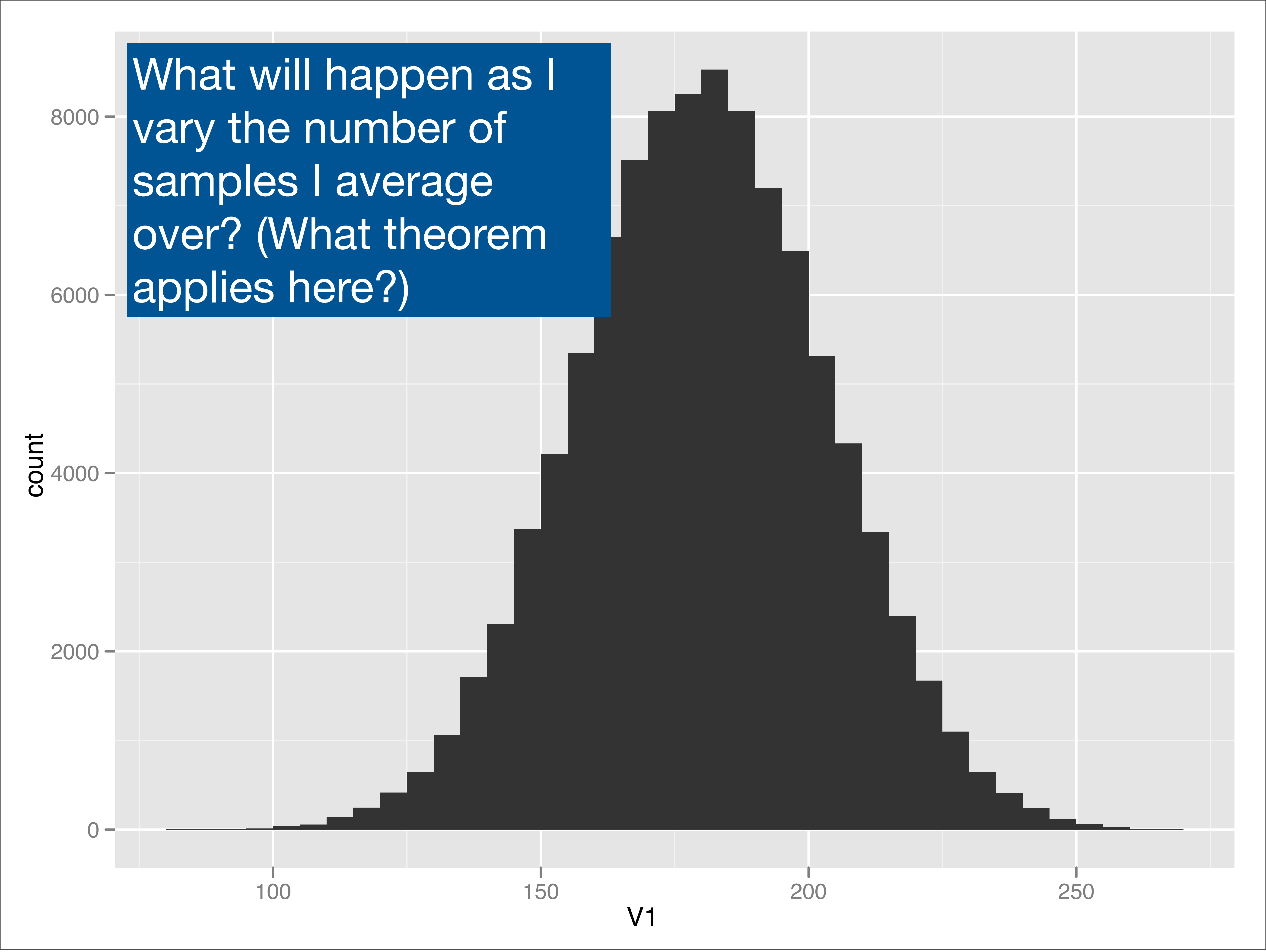


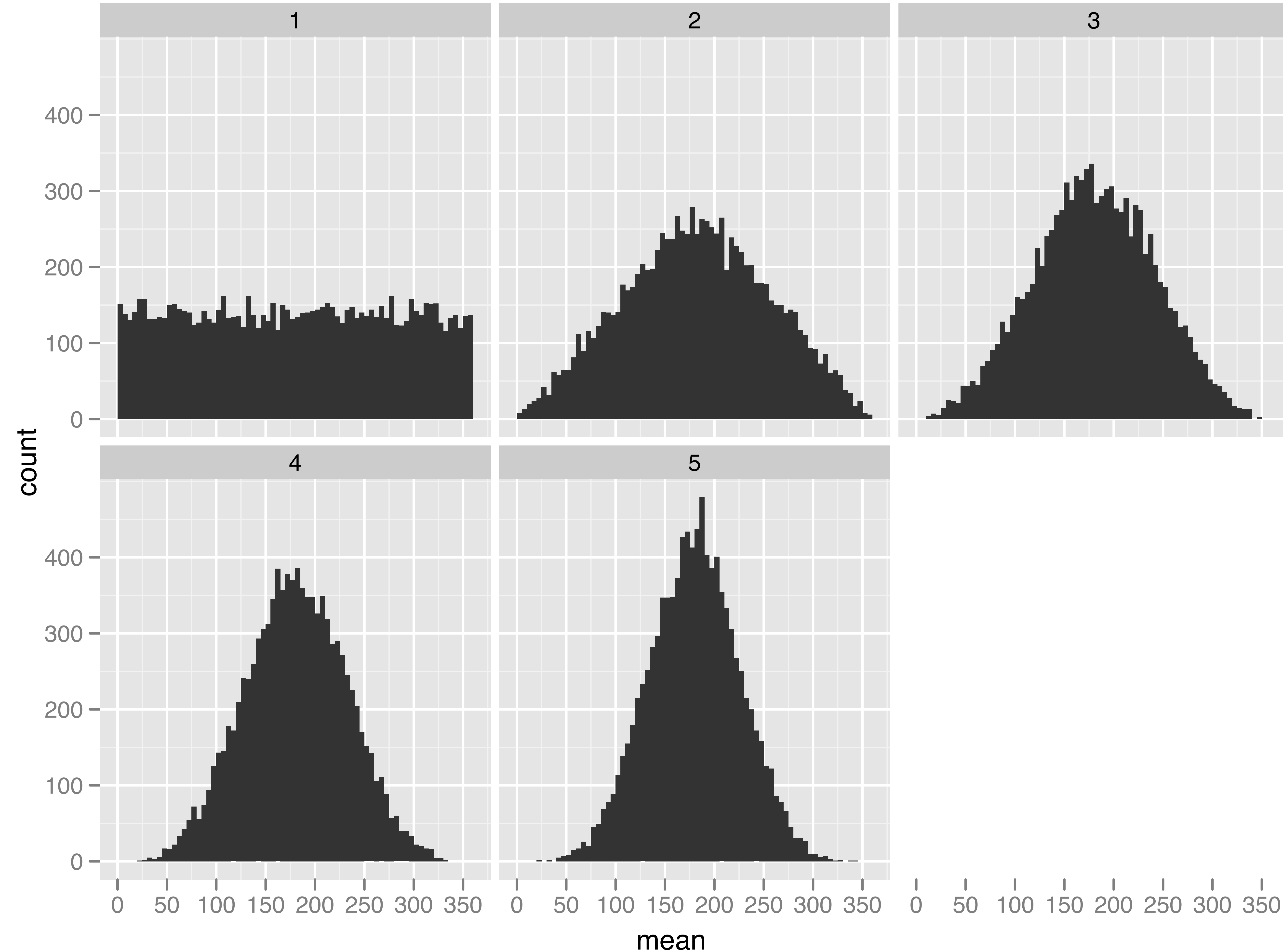
Experiment
Value
20
15
10
5
50
100
150
200
250
300
350

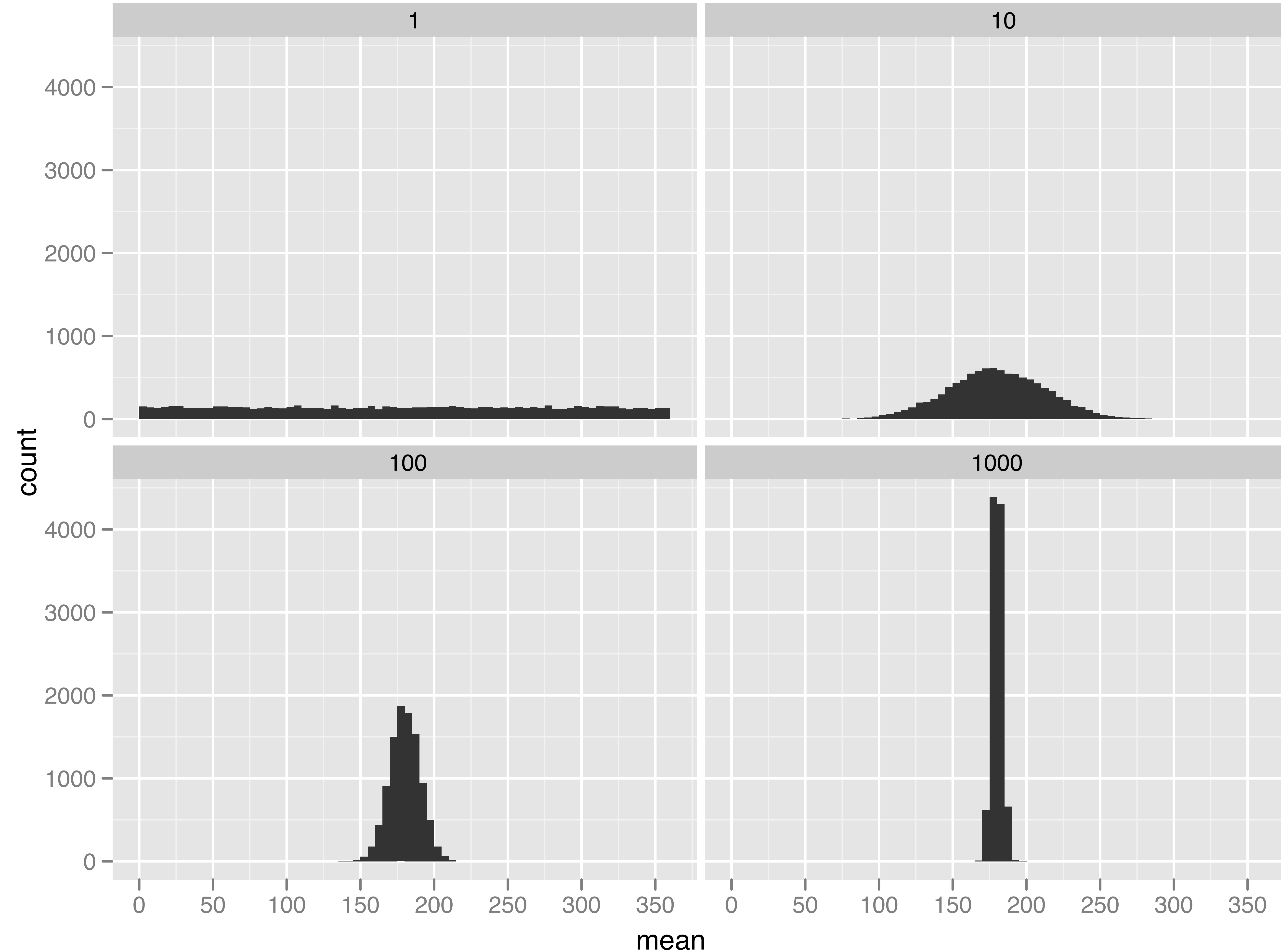


count
4
3
2
1
0
140
160
180
200
samp

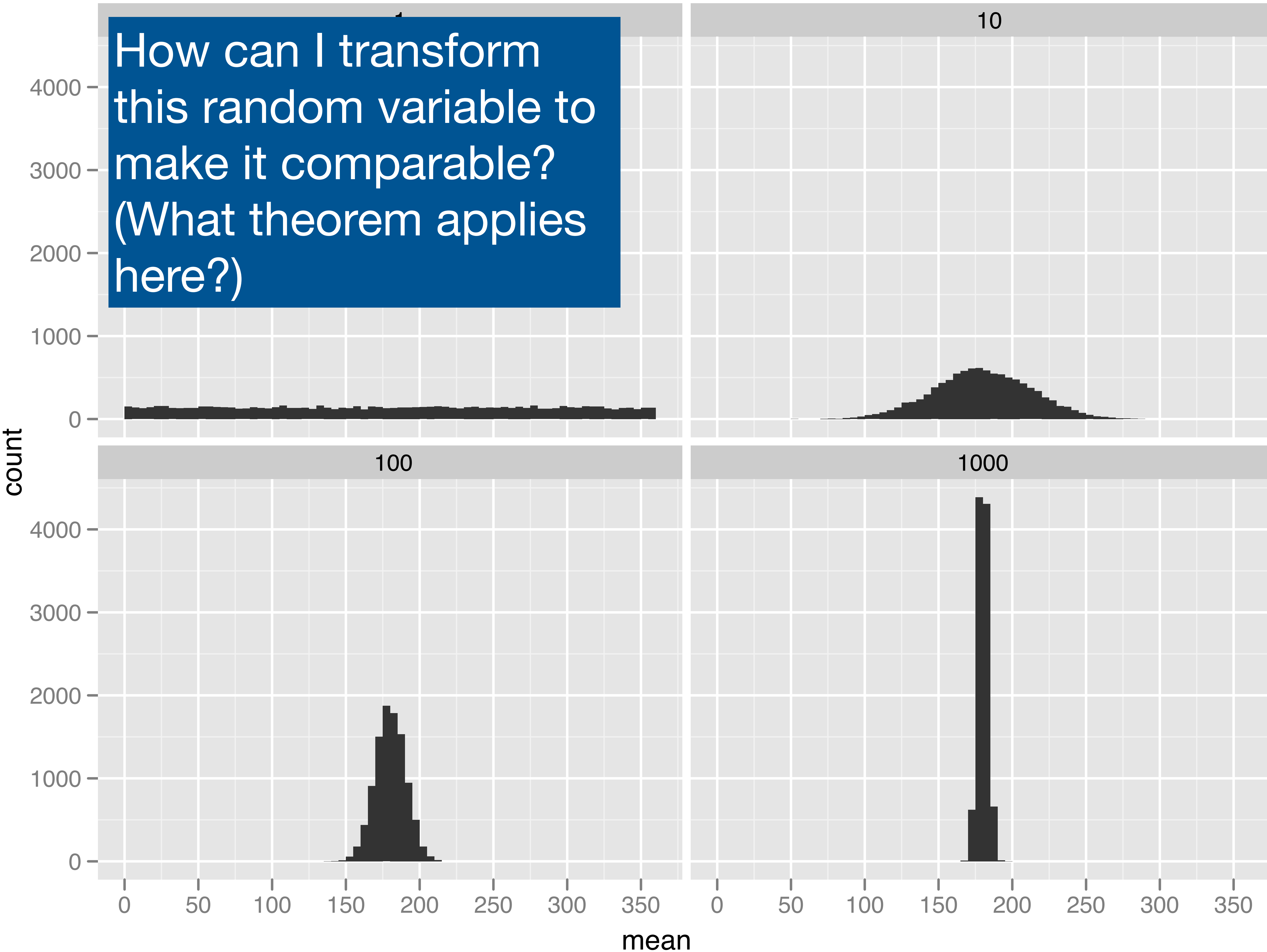

count
8000
6000
4000
2000
0
100
150
200
250
V1

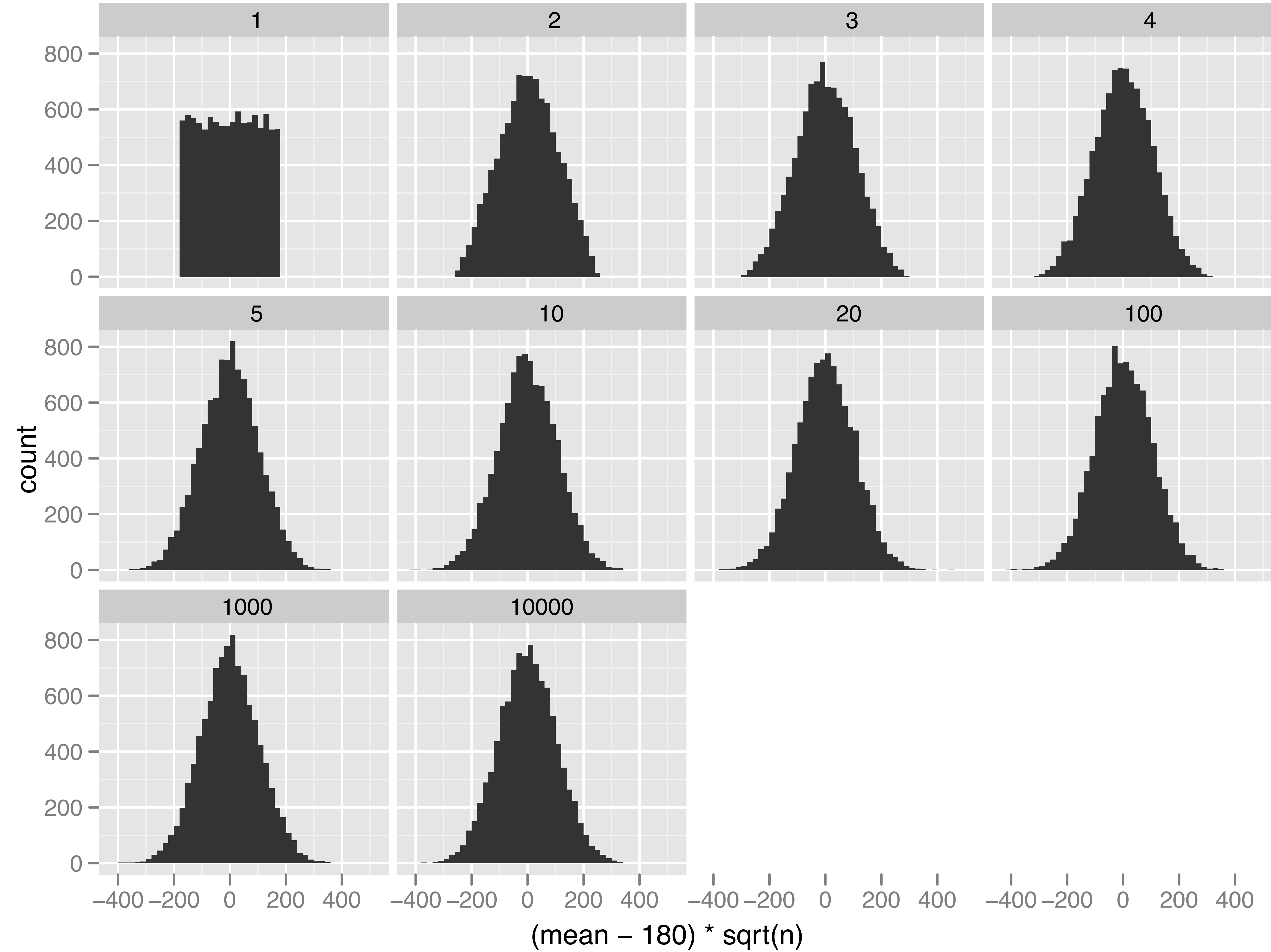

What will happen as I vary the number of samples I average over? (What theorem applies here?)
8000
6000
4000
2000
0
count
100
150
200
250
V1


1
2
3
4
5
count
mean
0
100
200
300
400
0
50
100
150
200
250
300
350

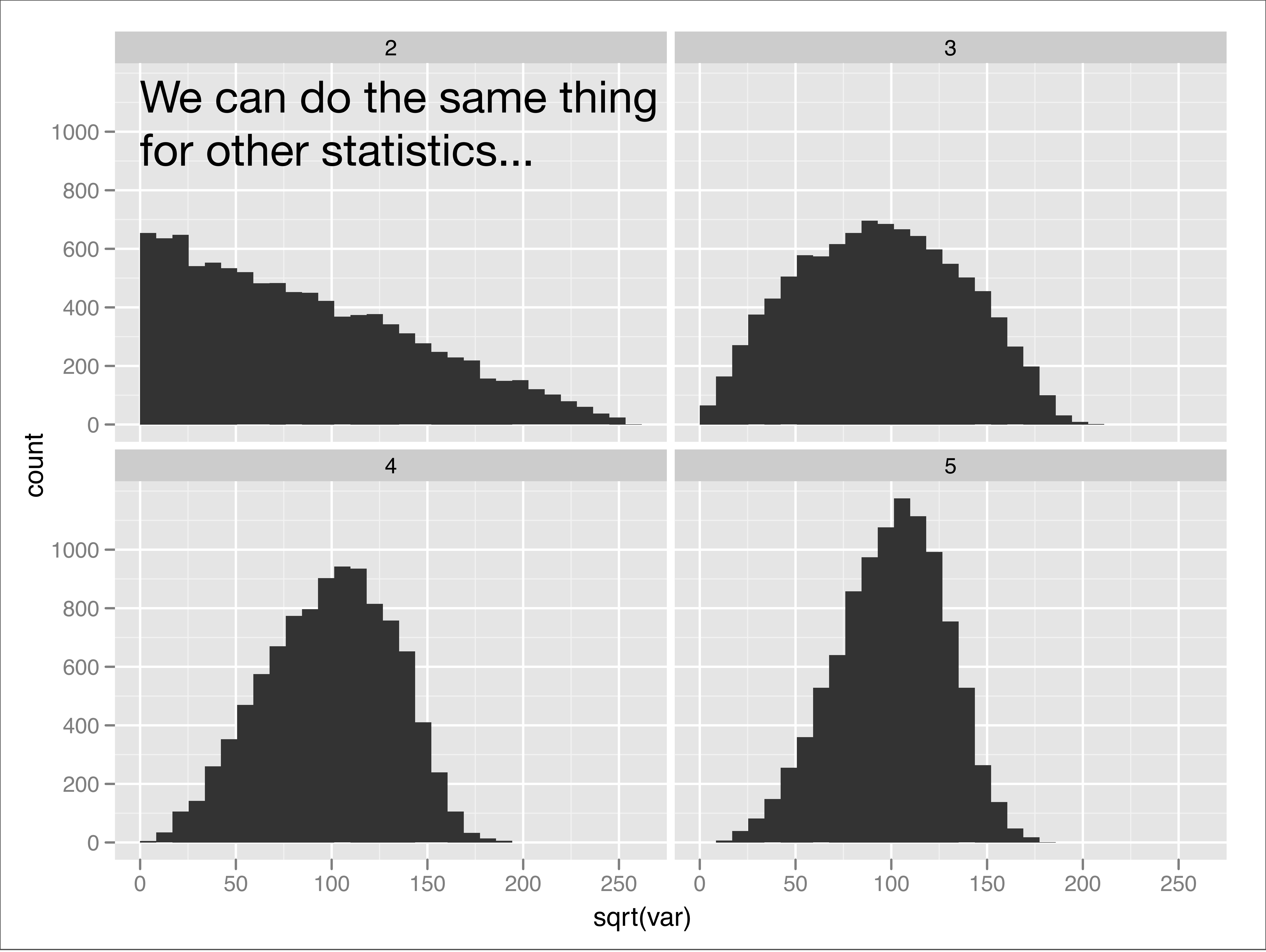

1
10
100
1000
count
0
1000
2000
3000
4000
0
50
100
150
200
250
300
350
mean

How can I transform this random variable to make it comparable? (What theorem applies here?)

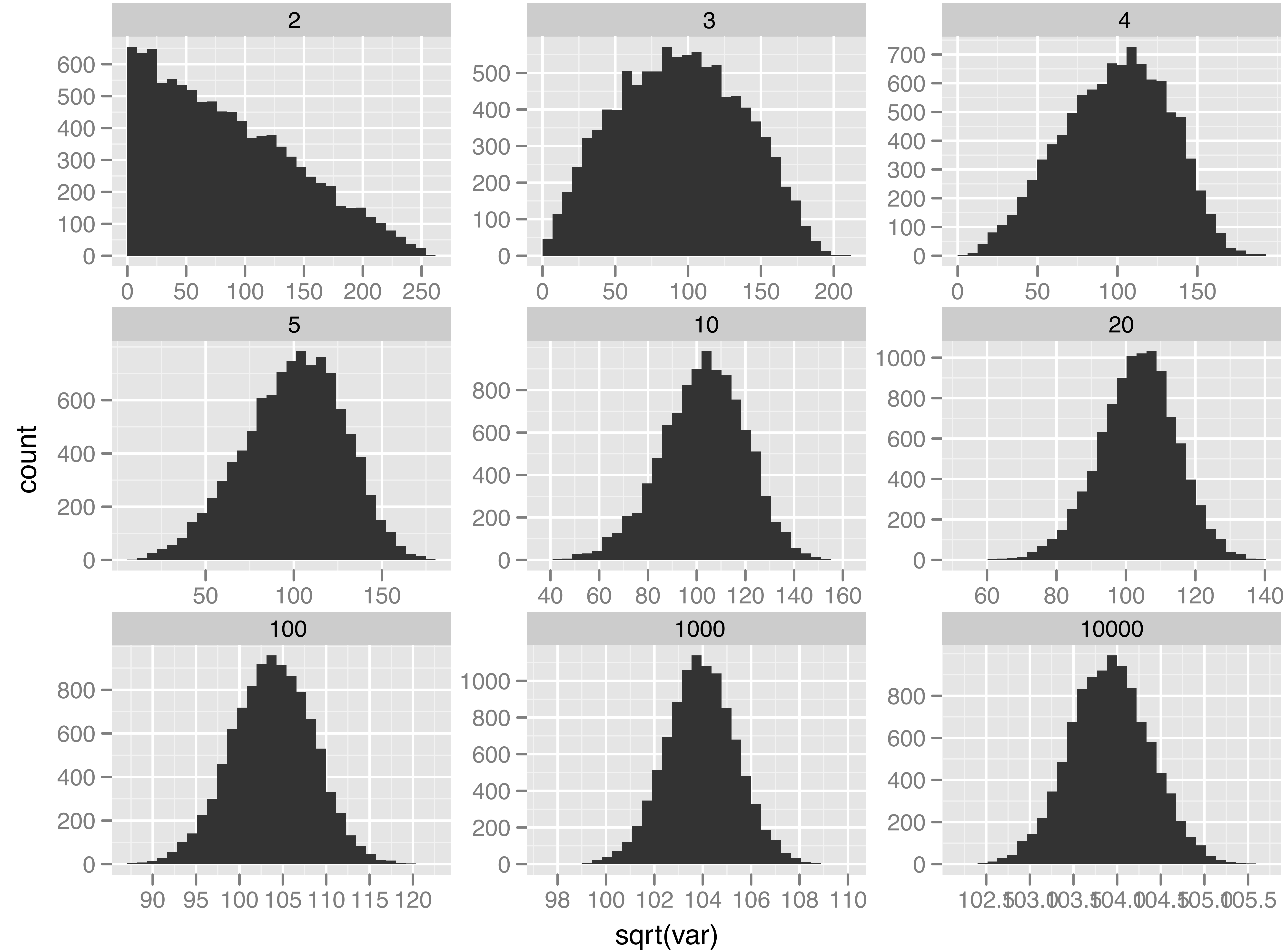




1
2
3
4
5
10
20
100
1000
10000
count
0
200
400
600
800
-400
-200
0
200
400
(mean – 180) * sqrt(n)



2
3
4
5
We can do the same thing
for other statistics...
count
1000
800
600
400
200
0
0
50
100
150
200
250
sqrt(var)


2
3
4
5
10
20
100
1000
10000
count
sqrt(var)

Next time

We'll start with the mean of normally distributed random variables, then try to extend in various ways.