Error Theory

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #11: Tuesday, February 11 2025

Announcements

Announcements

- No class the week of Feb. 25–27
  - Discuss scheduling makeup classes

Learning Goals

Learning Goals

- Limit theorems for binomial and Poisson distributions
- Using Q-Q plots to test for normality
- Basic error theory:
  - Different kinds of errors
  - Difference between *precision* and *accuracy*
  - Difference between *absolute* and *relative* errors
  - Difference between *random* and *systematic* errors
  - Correlated errors
  - Propagation of errors
- Testing the Central Limit Theorem

Additional Limit Theorems

Additional Limit Theorems

- Binomial $\longrightarrow$ Poisson:

$$X \sim \mathcal{B}(n, p)$$
$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - $\binom{n}{k}$ becomes hard to calculate when n is large.
  - For large n and small p, the binomial distribution approaches a Poisson distribution with $\lambda = np$

$$\mathbb{P}(X = k) \to e^{-\lambda} \frac{\lambda^k}{k!}$$

- Poisson $\longrightarrow$ Normal
  - This is slightly different to what was presented in the book.
  - As λ gets large, the Poisson distribution approaches a normal distribution with $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$

Set Up R Session

Accept GitHub Classroom Project

- Go to the GitHub Classroom Project at https://classroom.github.com/a/HNhvi1g2

- Create a new RStudio project using version control, from the GitHub Classroom assignment.

Set Up R Session

- Set up parameters and variables:

```
library(tidyverse)

set.seed(34593)

N <- 30
n_rep <- 500
k <- 2
theta <- 5
```

- Draw 500 replicates, each containing 30 samples from a gamma distribution

```
x <- map(1:n_rep, \(x) rgamma(N, shape = k, scale =
    theta))
```

- Calculate averages of each sample

```
x_bar <- map_dbl(x, mean)
```

Identifying Distributions

Identifying Distributions

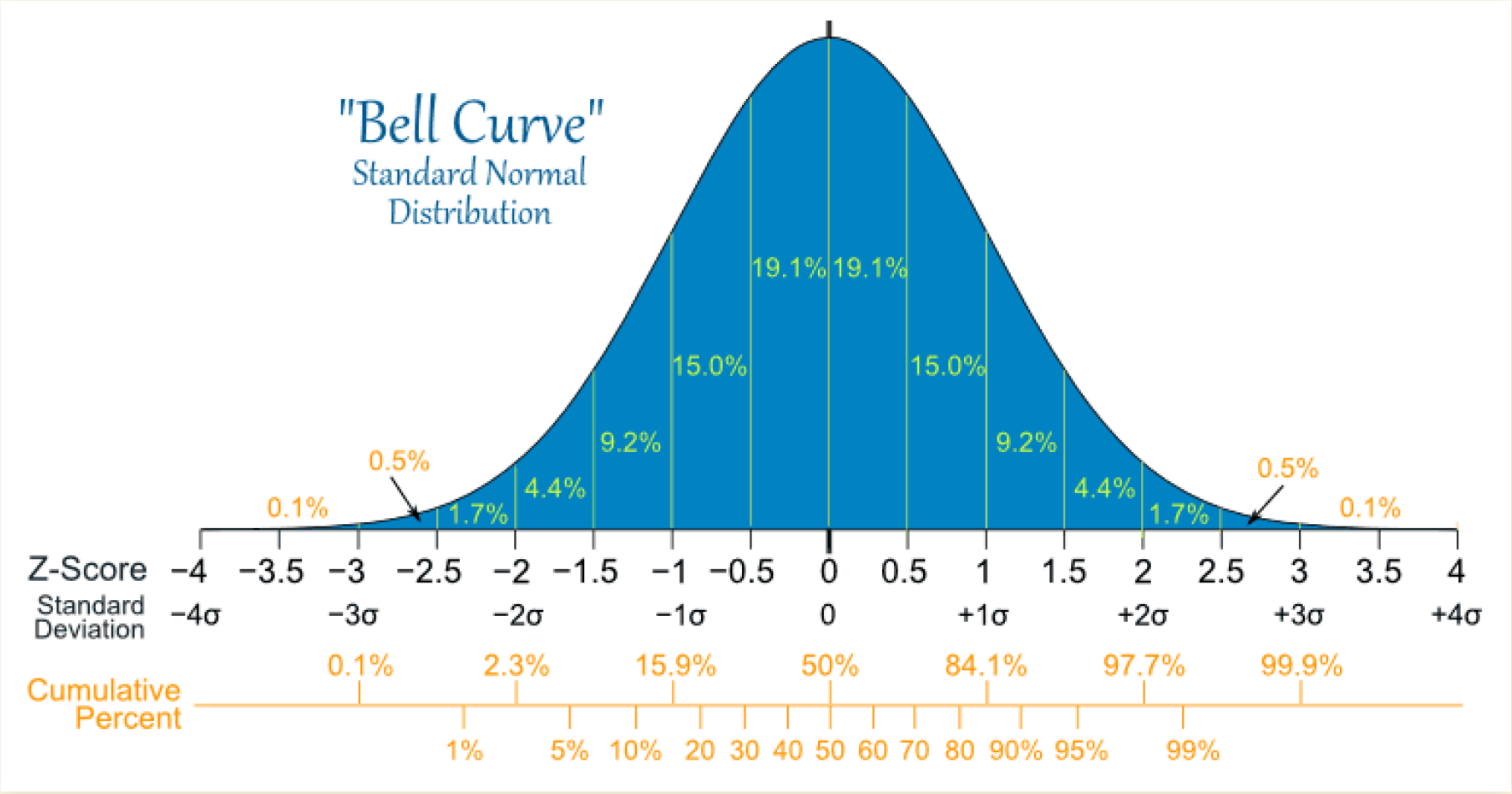
- How can you tell whether a sample of data is normally distributed?
  - Numerical tests, like Shapiro-Wilk test.
    - We won't use these
  - Graphical tests: Q-Q plots
    - Quantile-Quantile
      - Sort your sample from smallest to largest
      - Standardize your sample (mean = 0, sd = 1)
      - Each point represents a quantile
        - If there are 10 points in the sample, they're *deciles*
      - Make a scatterplot of the quantiles in your sample vs. quantiles for a normal distribution

Q-Q Plots

- Data: N points
- Sort the sample from smallest to largest
  - $x_1, x_2, \ldots, x_N$
- Standardize the sample:

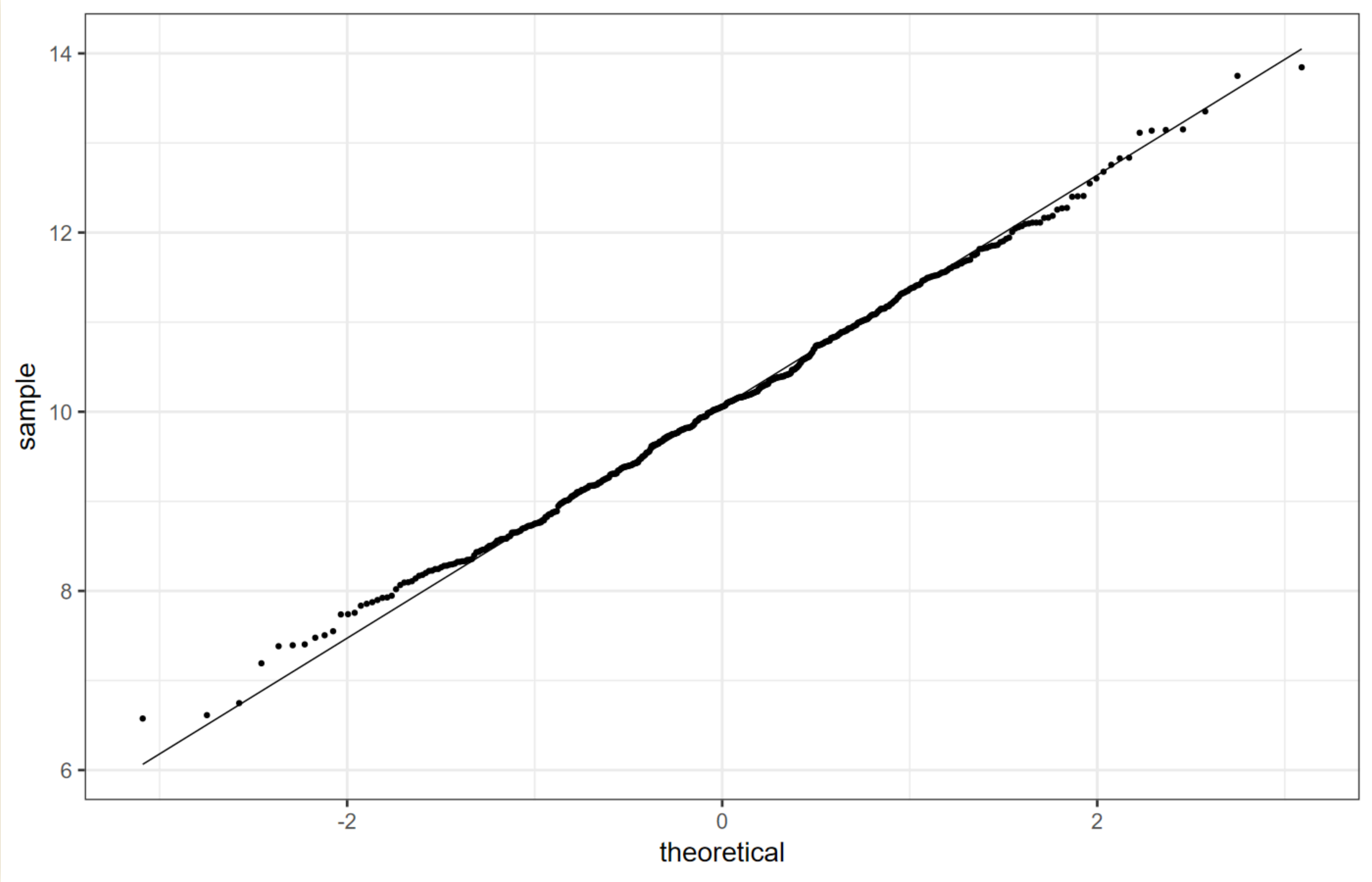
$$z_i = \frac{x_i - \overline{x}}{\mathrm{sd}(x)}$$

- Calculate $w_1, w_2, \ldots, w_N$, the N quantiles of the normal distribution



- Make a scatterplot (z_i, w_i)
- `geom_qq()` does all of this automatically.

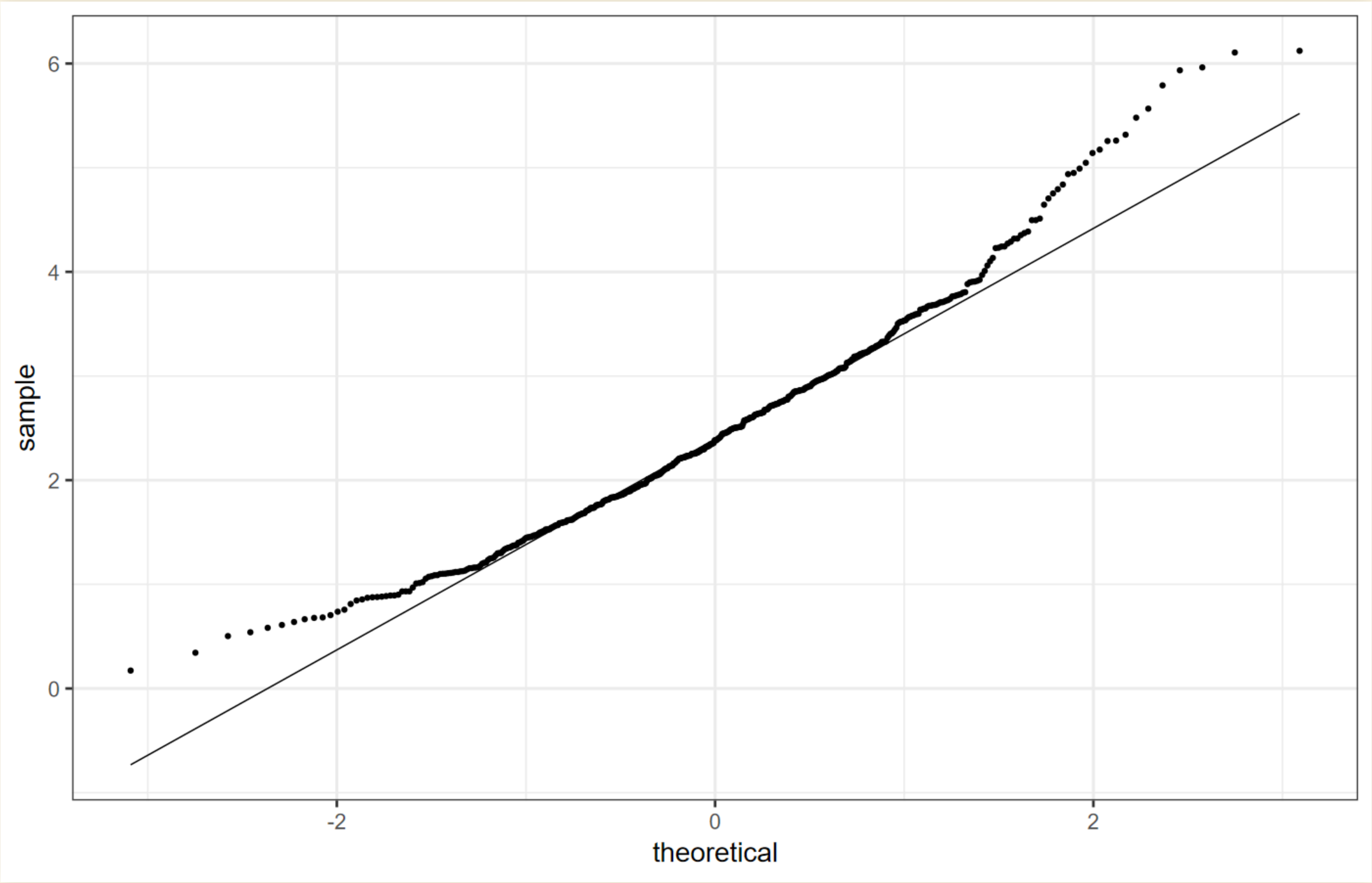
```
df <- tibble(x_bar = x_bar)
ggplot(df, aes(sample = x_bar)) +
  geom_qq_line() + geom_qq()
```



Q-Q Plots

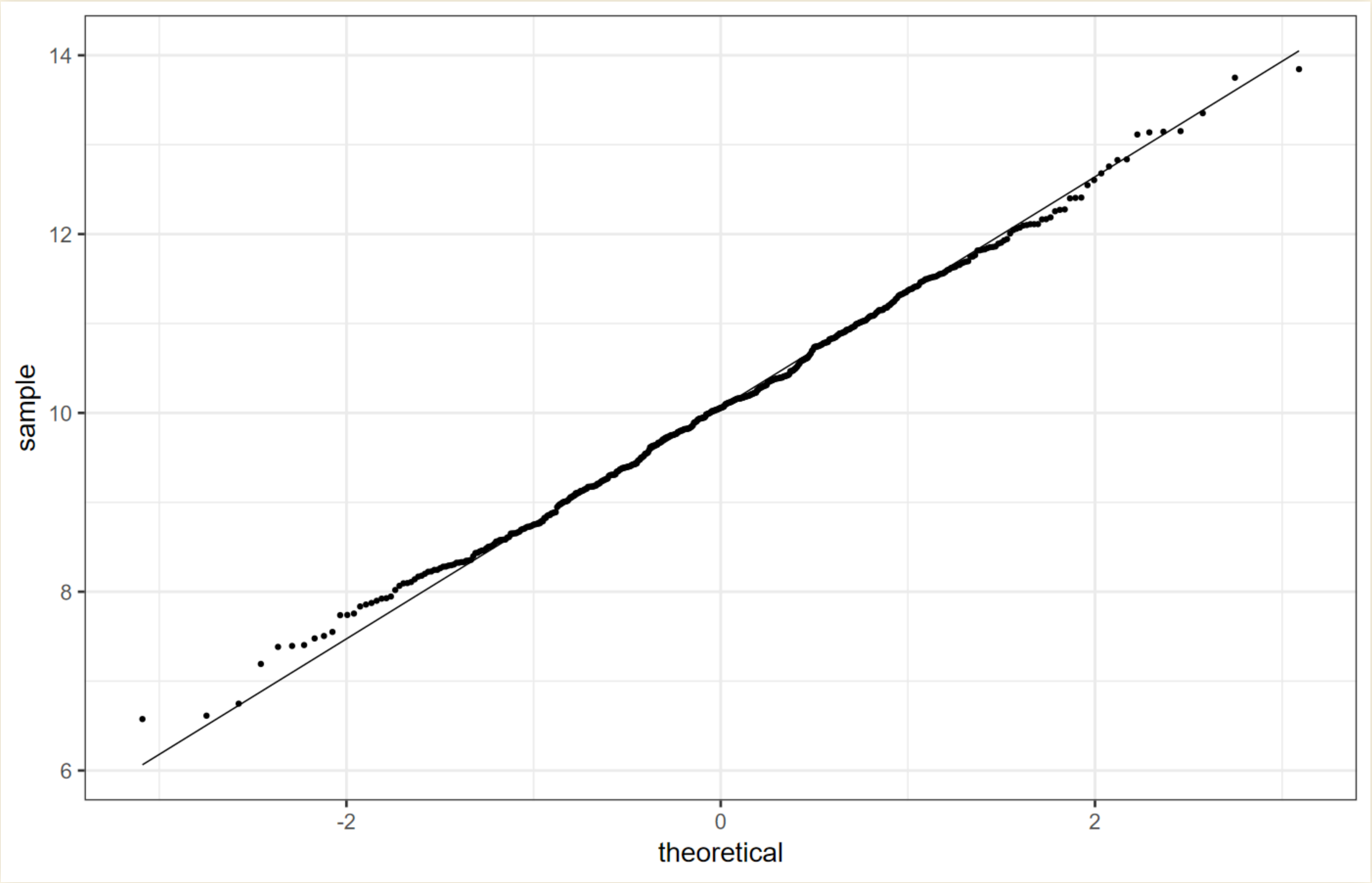
- 500 samples from a gamma distribution

```
df_gamma = tibble(x = rgamma(n_rep, k, shape = theta))
ggplot(df_gamma, aes(sample = x)) +
  geom_qq_line() + geom_qq()
```



- Means of 500 replications of 30 samples from a gamma distribution

```
ggplot(df, aes(sample = x_bar)) +
  geom_qq_line() + geom_qq()
```

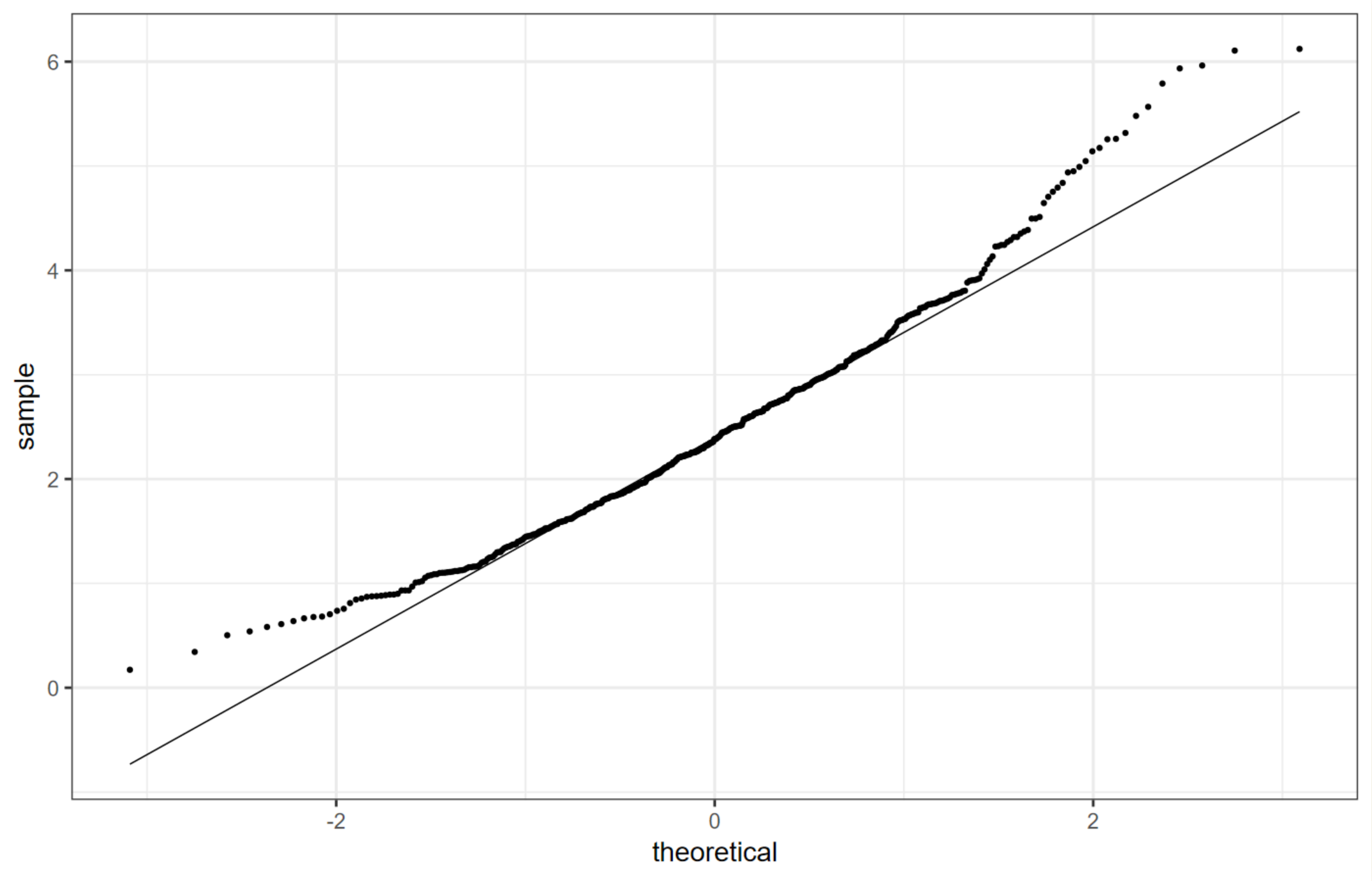


Interpreting Q-Q Plots

- **sample** $>$ **theoretical** at lower (left) end: The lower tail of the sample is narrower than the normal
- **sample** $>$ **theoretical** at upper (left) end: The upper tail of the sample is longer than the normal
- The sample is skewed:
  - Compared to a normal distribuiton:
    - You're *less* likely to find samples *far below* the mean
    - You're *more* likely to find sample *far above* the mean

- 500 samples from a gamma distribution

```
ggplot(df_gamma, aes(sample = x)) +
  geom_qq_line() + geom_qq()
```



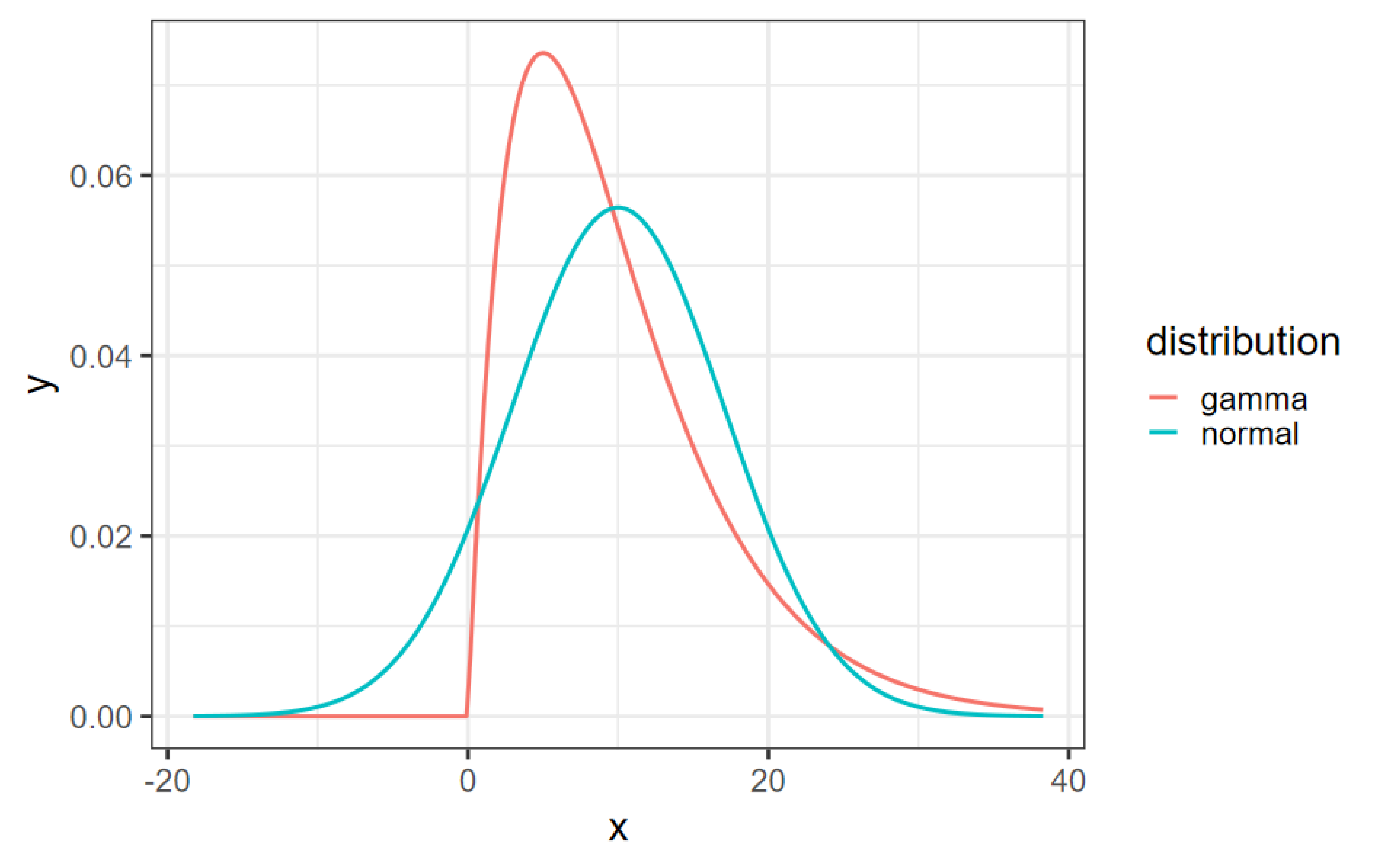
Comparing the sample to normal

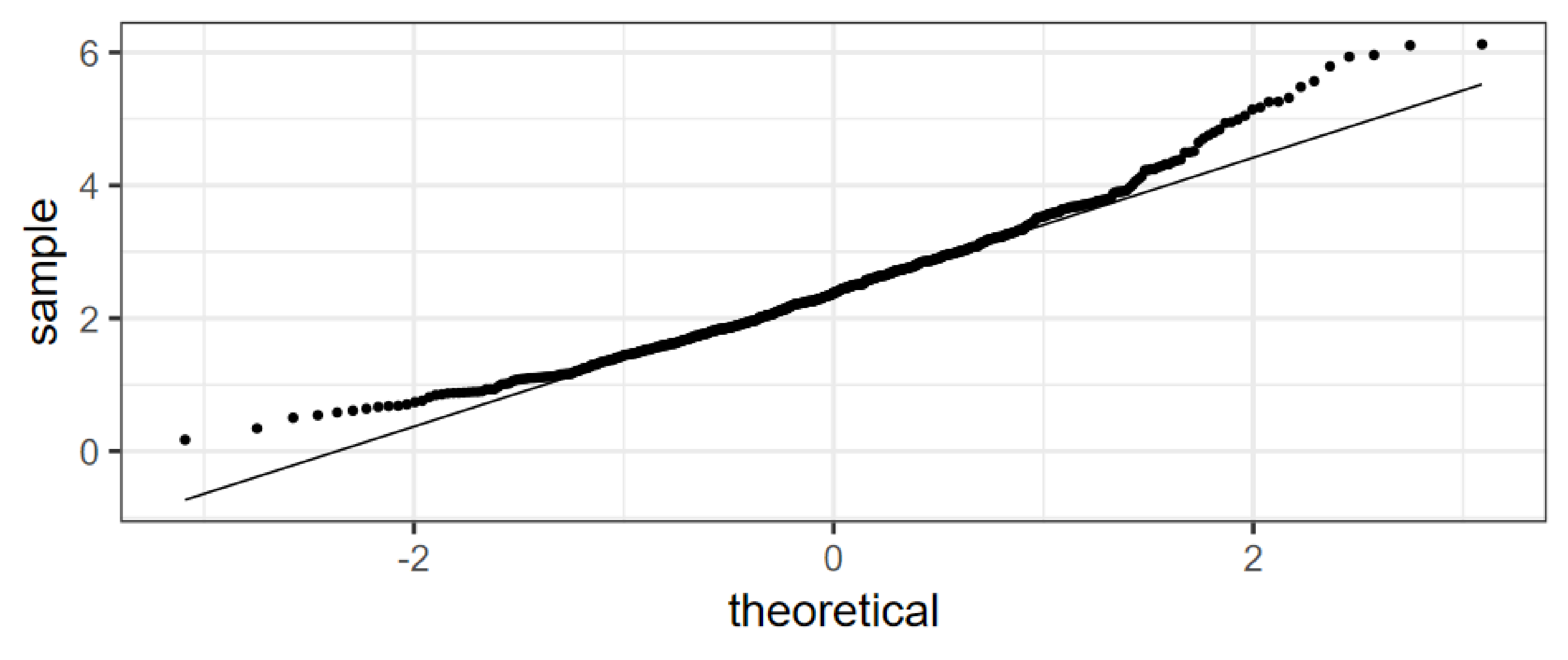
- The mean of a gamma distribution is $k \times \theta$ and the standard deviation is $\sqrt{k} \times \theta$:

```
mu <- k * theta
sigma <- sqrt(k) * theta
spread <- 4 * sigma

df_norm_gam <- tibble(
  x = seq(mu - spread, mu + spread, length.out = 200),
  gamma = dgamma(x, shape = k, scale = theta),
  normal = dnorm(x, mu, sigma)
) |>
  pivot_longer(gamma:normal, names_to = "distribution",
               values_to = "y")
ggplot(df_norm_gam, aes(x = x, y = y, color =
         distribution)) +
  geom_line(size = 1)
```

```
ggplot(df_gamma, aes(sample = x)) +
  geom_qq_line() + geom_qq()
```





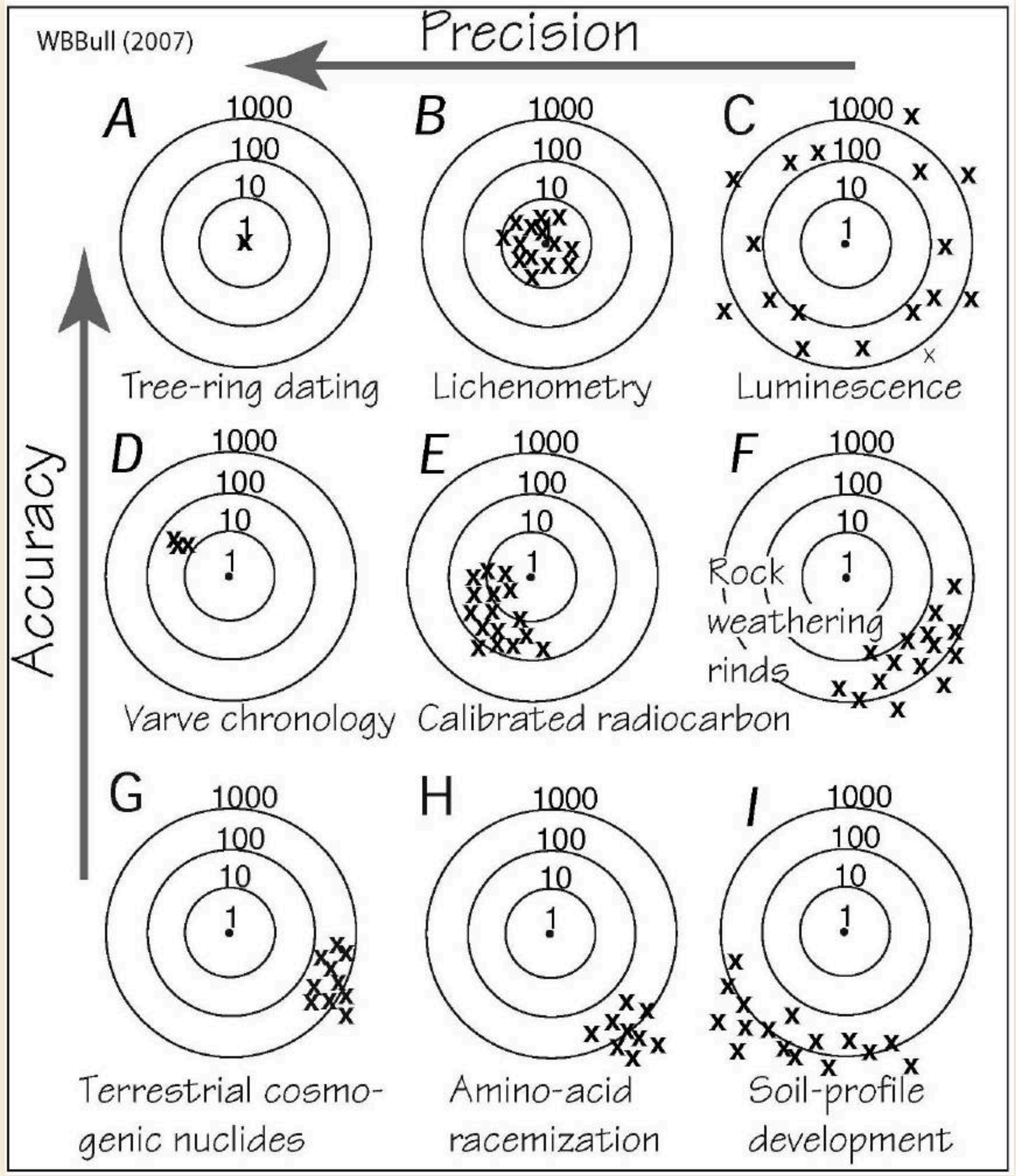
Error Theory

Error Terminology

- **Accuracy:** How far is the *mean* of your measurements from the true value?
- **Precision:** How close are your measurements to one another (how small is the *variance* of your measurements)?
- **Absolute error:** $\Delta x_{\text{abs}} = x - x_{\text{true}}$
- **Relative error:**

$$\Delta x_{\text{rel}} = \frac{x - x_{\text{true}}}{x}$$

- **Random error:** Errors are randomly distributed (usually normal)
- **Systematic error** or **bias:** Every measurement has the same error
- If *bias* is small:
  - More measurements $\longrightarrow$ better accuracy
  - Better precision $\longrightarrow$ better accuracy



Modeling Errors

- Common Model:

$$\text{Measurement} = ``\text{true value}'' + \text{error}$$
$$x = \mu + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma)$.

  - Errors are not always normally distributed
    - This is where the central limit theorem helps:
      - Even if errors of individual measurements are not normal, the average of many errors *will* be normal

- Covariance & Correlated Errors
  - You measure two variables, X, and Y
    - (e.g., temperature and humidity)
  - **Covariance:**

$$\text{Cov}(X, Y) = E(\,(X - E(X))\,(Y - E(Y))$$
$$\text{Cov}(X, X) = E(\,(X - E(X))\,(X - E(X))$$
$$= E(\,(X - E(X))^2\,)$$

  - If X and Y are independent, then $\text{Cov}(X, Y) = 0$.
  - If $\text{Cov}(X, Y) = 0$, X and Y *might not be* independent!
    - But if X and Y are *normal*, then $\text{Cov} = 0$ implies independence

Error Propagation

- If we measure X and Y with errors ΔX and ΔY: $X \pm \Delta X$ and $Y \pm \Delta Y$,
  - where ΔX and ΔY are *independent*, then
  - the error on $X + Y$ and $X - Y$ will both be

$$\Delta(X + Y) = \Delta(X - y) = \sqrt{(\Delta X)^2 (\Delta Y)^2}$$

  - the error on $X \times Y$ will be

$$\frac{\Delta(XY)}{XY} \approx \frac{\Delta X}{X} + \frac{\Delta Y}{Y}$$

and

$$\frac{\Delta(X/Y)}{X/Y} \approx \frac{\Delta X}{X} + \frac{\Delta Y}{Y}$$

Exploring the Central Limit Theorem

Exploring the Central Limit Theorem

- How many samples do you need for the mean of the samples to be normally distributed?
- Explore this with R
  - Repeat an analysis for different values of N
  - Use R functions to keep things organized

RStudio Project

- Open `test-central-limit-theorem.qmd`