Statistical Testing

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #15: Tuesday, March 04 2025

Announcements

Announcements

- Thursday will be review: I will work examples of the tests we’re learning about today.

Learning Goals

Learning Goals

- Review the $t - test$: understand p-values and confidence intervals
- Review the logic of statistical tests
- Understand the kinds of errors you can make in statistics:
- Understand *Test Power* and how we use this to select statistical tests and design experiments
- Learn about common parameteric and non-parametric tests
- Learn about how to test goodness of fit

Student's t-test

Two Tribes:

- Two tribes lived in an area for more than 1000 years
- Some expert archaeologists think
  - Tribe A arrived in 622 CE
  - Tribe B arrived in 615 CE
- Others dispute this.
- Archaeologists ask you to use ^{14}C dating to estimate ages of wood artifacts from early settlements of both tribes
  - The results of your measurements are:
    - Tribe A: $\overline{t_A} = 650\ \text{CE} \pm 50\text{y}\ (1\sigma)$
    - Tribe B: $\overline{t_B} = 750\ \text{CE} \pm 50\text{y}\ (1\sigma)$

One-Sample t-test

- Null hypothesis H_0: $\mu_A \leq 622$
- Alternate hypothesis H_a: $\mu_A > 622$
- One-sided, one-sample t-test:
  - T-statistic

$$\hat{T} = \frac{\overline{t_A} - \mu_A}{S_A/\sqrt{n_A}}$$

  - Compute $\mathbb{P}(t > \hat{T}) = 1 - F_{t_\nu}(\hat{T})$, where F_{t_ν} is the cumulative distribution function of the t-distribution for ν degrees of freedom.

- Suppose $\hat{T} = 1.9$.
  - If $n_A = 4, 1 - F_{t_3}(1.9) = 8\%$, so we can't reject H_0 at the 5% level.
  - If $n_A = 12, 1 - F_{t_{11}}(1.9) = 4\%$, so we reject H_0 at the 5% level.
- 4 measurements aren't enough to tell the difference between tribe A arriving before or after 622 CE.
- 12 measurements are sufficient to tell the difference, and confidently say that the tribe probably arrived after 622.

One-Sample *t*-Test in R

- Sample some data:

```
set.seed(179011)
x_A4 <- rnorm(4, 650, 50)
x_A12 <- rnorm(12, 650, 50)
```

  - Run a t-test

```
t.test(x_A4, mu = 622, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  x_A4
## t = -1.8837, df = 3, p-value = 0.9219
## alternative hypothesis: true mean is greater than
622
## 95 percent confidence interval:
##  587.5994      Inf
## sample estimates:
## mean of x
##  606.7063
```

- Now try with 12 samples

```
t.test(x_A12, mu = 622, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  x_A12
## t = 4.4402, df = 11, p-value = 0.0004974
## alternative hypothesis: true mean is greater than
622
## 95 percent confidence interval:
##  650.0727      Inf
## sample estimates:
## mean of x
##  669.1384
```

  - 4 samples: $\hat{T} = -1.9$, $p = 0.92$, so we can't reject H_0.
    - 4 samples isn't enough to tell whether tribe A arrived before or after 622 CE.
  - 12 samples: $\hat{T} = 4.4$, $p = 5 \times 10^{-4}$, so we can confidently reject H_0
    - With 12 samples we can confidently tell that tribe A arrived after 622 CE.

Two-Sample t-Test

- Null hypothesis H_0: $\mu_B \leq \mu_A$
- Alternate hypothesis H_a: $\mu_B > \mu_A$
- One-sided two-sample t-test:
  - Compute the two-sample T-statistic

$$\hat{T} = \frac{\overline{t_B} - \overline{t_A}}{\sqrt{\frac{S_B^2}{n_B} + \frac{S_A^2}{n_A}}} \sim t_{\nu'}$$

    where t_ν' is the student-t distribution and ν' depends on what we know about whether t_A and t_B have the same variance.
    - This equation means that the T statistic $\hat{T}$ behaves like a random variable drawn from the $t_{\nu'}$ distribution.
  - R will calculate ν' so we don't have to worry about the formulas in the textbook

- Try it in R

```
x_B <- rnorm(9, 750, 50)
t.test(x_B, x_A12, alternative = "greater", var.equal =
        TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x_B and x_A12
## t = 6.8518, df = 19, p-value = 7.715e-07
## alternative hypothesis: true difference in means is
greater than 0
## 95 percent confidence interval:
##  71.08198      Inf
## sample estimates:
## mean of x mean of y
##  764.2136  669.1384
```

- The p-value is $7.7{\times}10^{-7}$, so we reject the H_0 because there is only a 0.00008% chance that we'd see this data if H_0 were true.
  - We conclude the tribe B arrived *after* tribe A ($\mu_B > \mu_A$)

The Logic of Statistical Tests

The Logic of Statistical Tests

- Five Steps:
  1. Identify the appropriate test and test statistic
     - e.g., t-test and T statistic
  2. Define the null hypothesis
     - e.g., H_0: $\mu_1 = \mu_2$
  3. Define an alternate hypothesis:
     - e.g., H_a: $\mu_1 > \mu_2$ (one-sided)
     - H_a: $\mu_1 \neq \mu_2$ (two-sided)
  4. Obtain the *null distribution*
     - Distribution of the test statistic if H_0 is true
  5. Compute p-value
     - Probability that you'd see values as extreme as the observed test statistic if H_0 is true
     - Compare to test level α
       - e.g., $\alpha = 0.05$
     - $p < \alpha$: Reject H_0 (guilty)
     - $p \geq \alpha$: Insufficient evidence to reject H_0 (not guilty $\neq$ innocent)

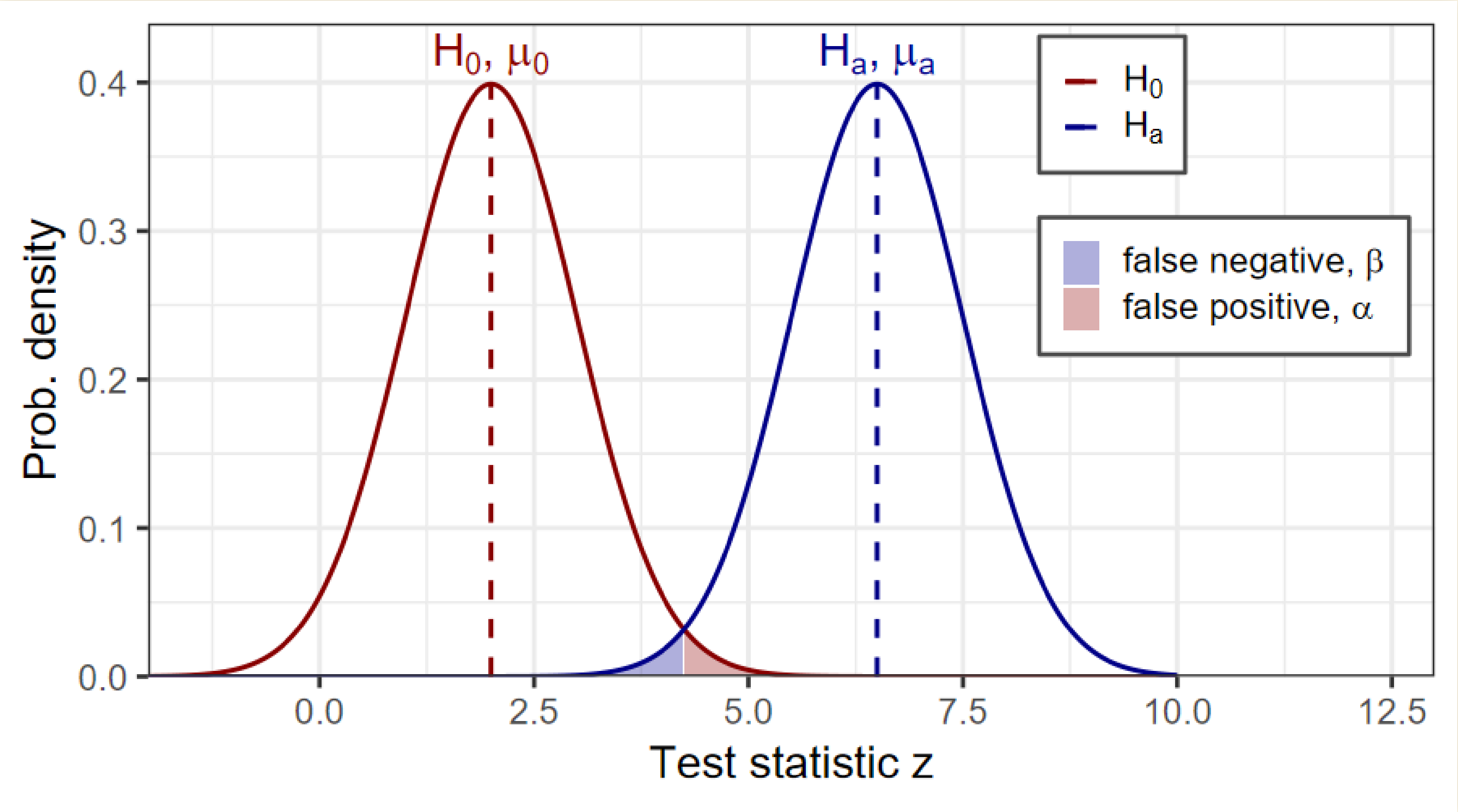
Test Errors

Test Errors

- Our statistical tests either reject H_0 or don't reject it.
  - Scenario: You're testing for COVID.
    - H_0: the patient doesn't have COVID.
    - H_a: the patient has COVID.
  - *Positive test result*: reject H_0.
    - Diagnosis: The patient has COVID.
  - *Negative test result*: don't reject H_0.
    - Diagnosis: The patient doesn't have COVID.
- Four possible outcomes:

Decision	**H_0 is true**	**H_0 is false**
Positive: Reject H_0	False positive	True positive
Negative: Don't reject H_0	True negative	False negative

- Correct results:
  - *true positive*: reject H_0 when it's false.
  - *true negative*: don't reject H_0 when it's true.
- Errors:
  - **Type-I error** (*false positive*, α): H_0 is true but we reject it.
  - **Type-II error** (*false negative*, β): H_0 is false, but we don't reject it.

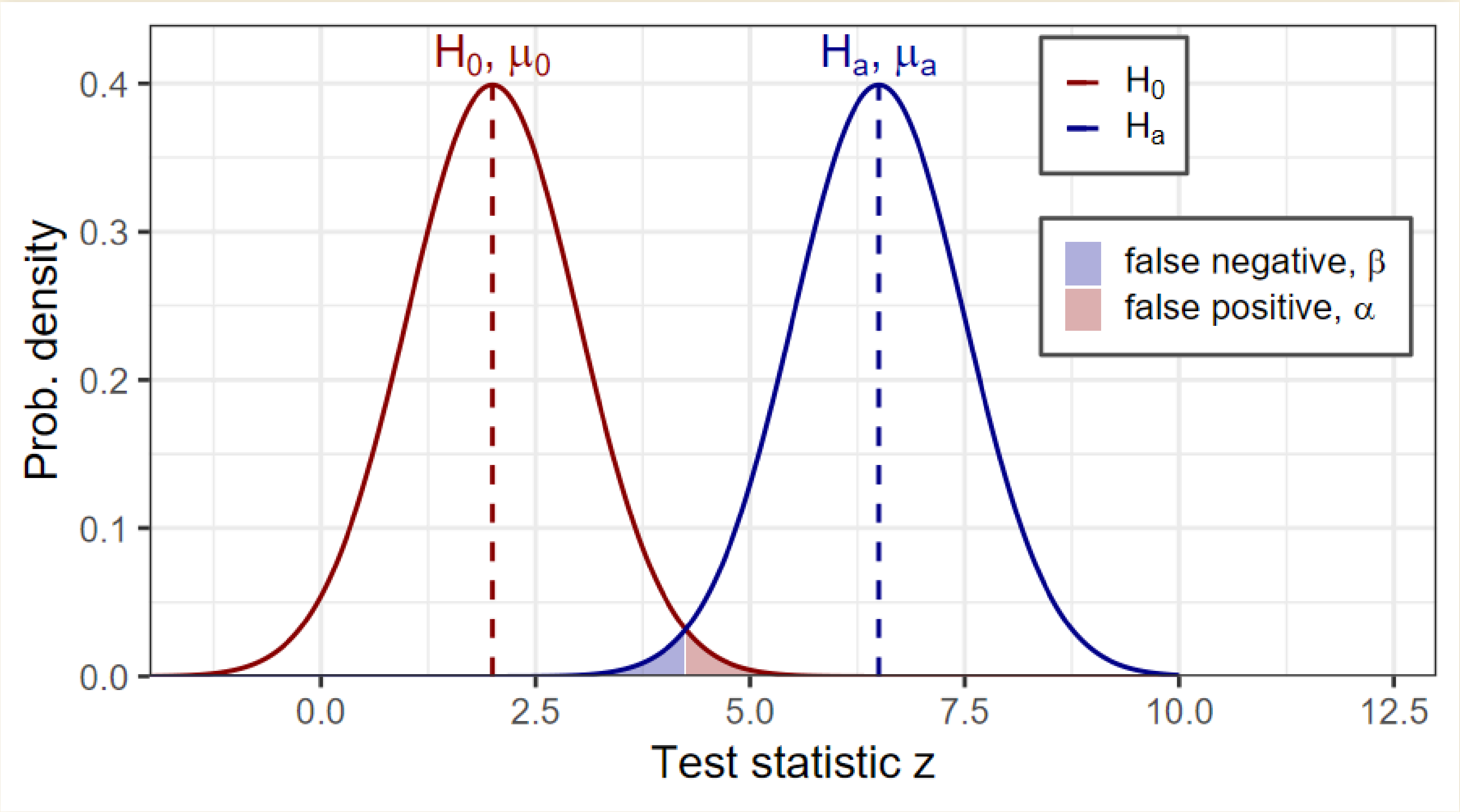


Test Power

- **Power**: Probability of rejecting H_0:

$$1 - \beta = \mathbb{P}(\text{rejecting } H_0 | H_a \text{ is true})$$

  - β is the probability of a *false negative* (Type II error) if H_a is true.
  - Power depends on N, the # of observations
  - Measures *discrimination*:
    - How well can a test discriminate between H_0 and H_a?



- We often use *power analysis* when designing an experiment to estimate how large a sample we need (how many observations) to detect an effect with confidence.

- When designing a statistical test, there's a tradeoff:
  - Making β smaller makes α larger and vice-versa.
  - We can use power analysis to choose which test to use, based on the tradeoff between α and β.

Statistical Tests

Parametric Tests

- If you know your data follow a *parametric distribution* (typically a *normal distribution* $\mathcal{N}$)
  - Z-test: Compares two means when the variance is known
  - t-test: Compares two means when the means and variances are unknown
  - F-test: Compares variances of two samples from two populations:

$$\begin{cases} \text{Sample 1, } \sigma_1 \text{, } n \text{ observations} \\ \text{Sample 2, } \sigma_2 \text{, } m \text{ observations} \end{cases} \qquad F_{m,n} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \frac{\chi^2_{n-1}}{\chi^2_{m-1}}$$

    - σ_1^2, σ_2^2 are the unknown true variances of the populations from which the samples are drawn.
    - S_1^2, S_2^2 are the observed variances of the samples.

Goodness of Fit

- How well does a theoretical distribution represent the observations?
  - H_0: the observations match the theoretical distribution
  - Tests determine whether you can reject H_0.
- χ^2 test
  - Compare histograms of observed and theoretical probability mass
  - If the fit is good, the # of observations O_k in each bin k should be close to the theoretical expectation E_k

$$\Xi^2 = \sum_{k=1}^{N_b} \frac{(E_k - O_k)^2}{E_k} \sim \chi^2_{\nu-1}(O, E),$$

  where $\nu = N_b - n_p$, N_b is the number of observations and n_p is the number of parameters you estimate to decribe the theoretical probabilty distribution.
    - Ξ behaves like a random variable drawn from a $\chi^2_{\nu-1}$ distribution.
  - In R, we use the `chisq.test()` function.

Kolmogorov-Smirnov Test

- Kolmogorov-Smirnov Test

$$D = \max_x |F_n(x) - F(x)|,$$

where $F_n(x)$ is the empirical cumulative distribution function for your observations and $F(x)$ is the theoretical cumulative probability distribution fucntion.

  - Measures the greatest discrepancy between empirical and theoretical cumulative distributions
  - Similar to measuring the greatest deviation from a straight line in a Q-Q plot.

- Reject H_0 at level α if

$$D > C_\alpha = \frac{k_\alpha}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}},$$

where n is the sample size and k_α is a function of α.

- The K-S test is universal: You can compare a sample of observations to any theoretical distribution
- Disadvantage: The test doesn't account for reducing the degrees of freedom when you use observations to estimate parameters in $F(x)$.
- In R, use `ks.test()`

Nonparametric Tests

Nonparametric Tests

- Most parametric tests only work if your data follow a parametric distribution
  - Most work only work for a Normal distribution
  - This is a historical artifact of what math people could do using pencil and paper
- Nonparametric tests substitute computer power for mathematical elegance
  - They work for any distribution
  - You don't need to know a mathematical formula for the distribution
- Basic strategy:
  - Simulate a large sample of *surrogate data* under the null hypothesis
  - Compare this sample to the observed data

- Four ideas that use Monte Carlo methods:
  1. Permutation
  2. Reordering
  3. Resampling
  4. Direct simulation

Sampling

- Important distinction:
  - Sample m numbered balls from a jar containing N balls.
- Sampling without replacement:
  - Draw balls without putting any back
    - Each ball can only appear once in the final sample
    - $m \leq N$
    - There are

$$\binom{N}{m} = \frac{N!}{m!(N-m)!}$$

      different ways to sample m balls.
    - $\binom{N}{m}$ is called the binomial coefficient
- Sampling with replacement:
  - Draw a ball, put it back, draw another, …
    - A ball may be drawn more than once
    - No limit to how big m can be
    - There are N^m different ways to sample m balls.

Permutation Tests

- Start with a sample of size $n = n_1 + n_2$
  - You want to compare the n_1 sample to the n_2 sample.
    - H_0: The two are the same
  - Generate N surrogate samples by sampling without replacement from the combined sample of size n.
    - This mixes up the two parts.
    - The number of possible samples in the surrogate ensemble is

$$N = \binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2} = \frac{(n_1 + n_2)!}{n_1!\,n_2!}$$

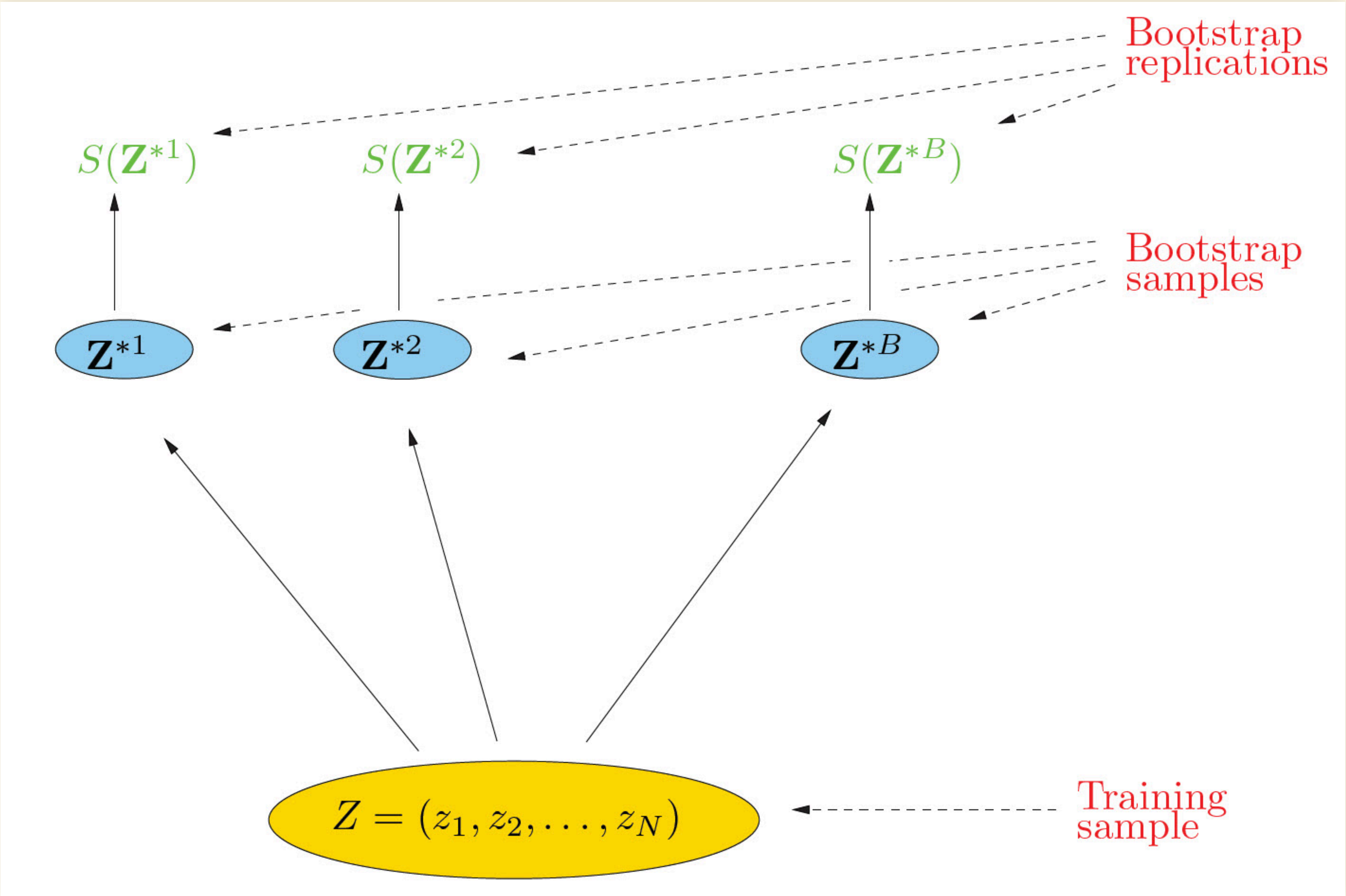
- Compare a test statistic Z on the original n_1 vs. n_2 parts to the distribution of test statistics from the ensemble of surrogate samples

Example Permutation Test

- Change in rainfall under global warming
  - You have 16 samples from climate model runs each of two scenarios: one with the preindustrial level of CO_2 ($1 \times CO_2$) and the other with double the preindustrial CO_2 ($2 \times CO_2$)
  - Choose a test statistic Z: The difference in maximum winter rainfall between the two groups of simulations
    - This is very far from Normally distributed
    - Parametric tests won't work
- Measure the test statistic comparing the two scenarios: Z_0
- Generate a large surrogate ensemble of random permutations $S_{1,k}$ and $S_{2,k}$ that sample from both scenarios and mix them up.
  - There are 6.0×10^8 possible permutations.
- Calculate test statistics Z_k from $S_{1,k}$ and $S_{2,k}$ for each surrogate permutation k.
- If Z_0 (the observed test statistic) seems very unlikely under the distribution of surrogate statistics Z_k, then you reject the null hypothesis and conclude that the rainfall is different between the two climate scenarios.

Resampling Tests: Bootstrap

- Bootstrap Sampling
  - Generate a large ensemble from a limited sample
    - Pull yourself up by your bootstraps
  - Start with sample $Z = (z_1, z_2, \ldots, z_N)$.
  - Generate B samples $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ each containing N values sampled from the original Z, with replacement.
    - There are N^N possible different samples
  - Calculate the test statistic $S(Z^{*i})$ for each Z^{*i}.
  - To get the confidence interval for confidence level α:
    - Sort the $S(Z^{*i})$ and pick the $B \times \alpha/2$ smallest and $B \times (1 - \alpha/2)$ largest values.



- How large should B be?
  - $B = 200$ is generally a good value.
- In R load `library(boot)` and use the function `boot()` or load `library(bootstrap)` and use the function `bootstrap()`

Resampling Tests: Jackknife

- Instead of drawing samples of N values from the original Z, make N samples of $N-1$ values, each of which leaves one value out (leave-one-out sampling).

$$\textsf{sample}_j = (z_1, z_2, \ldots, z_{j-1}, z_{j+1}, z_{j+2}, \ldots, z_N)$$

- These days, the bootstrap is much better than the jackknife, but there are other useful applications of leave-one-out sampling.

Direct Simulation Tests

- Sometimes you have a specific model of the process that generated the data.
  - Example: In time-series data, you may have autocorrelations, where the value of the next point depends on the values of one or more previous points.
    - This violates the IID assumption (that each sample is independent of the others, and drawn from an identical probability distribution).
  - We analyze these situations by using the computer to directly simulate the process that generated the data.
- We'll examine simulation methods when we study time-series data.