ST 516: Foundations of Data Analytics

Validity of Inference

We say that an inference procedure (method of constructing confidence intervals, performing hypothesis tests, or computing p-values) is **valid** if we can trust the probability statements that the procedure was designed to satisfy.

For instance, if we construct a 95% confidence interval, can we actually believe that 95% of the times that we construct such an interval, it will contain the true parameter value?

We should not base scientific decisions on the results of inference procedures that are very far from valid (that is, the probability statements are way off). Here we discuss in more detail what it means for inference to be valid.

Validity of a Statistical Procedure

- Hypothesis testing:
  A hypothesis testing procedure is **valid** if the actual probability of rejecting a true null hypothesis is the desired level α.

- p-values:
  A procedure for obtaining p-values is **valid** if the resulting value actually reflects the probability of obtaining results at least as extreme as the observed value, when the null hypothesis is true.

- Confidence Intervals: A $(1-\alpha)100\%$ confidence interval is **valid** if it contains the true parameter value $(1-\alpha)100\%$ of the time.

Validity of Confidence Intervals

Validity of Confidence Intervals:

A $(1-\alpha)100\%$ confidence interval is **valid** if it contains the true parameter value $(1-\alpha)100\%$ of the time.

To assess validity of confidence intervals, we can:

- Use known mathematical theory... or
- Simulate data with a known true parameter value, and examine how frequently the resulting confidence interval contains ('*covers*') the true value

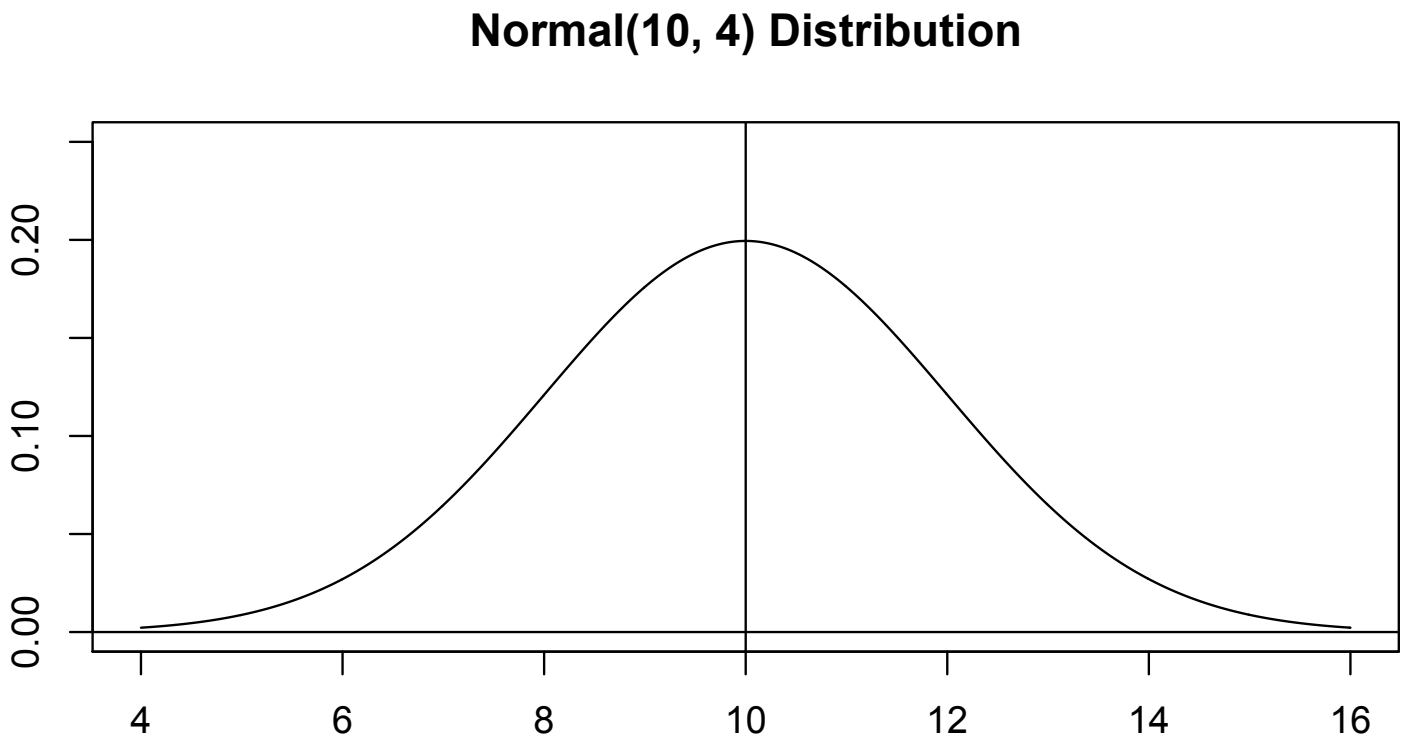
This allows us to explore how a particular interval construction *method* works in different settings:

- Different population distributions
- Different sample sizes

Validity of Confidence Intervals

Example

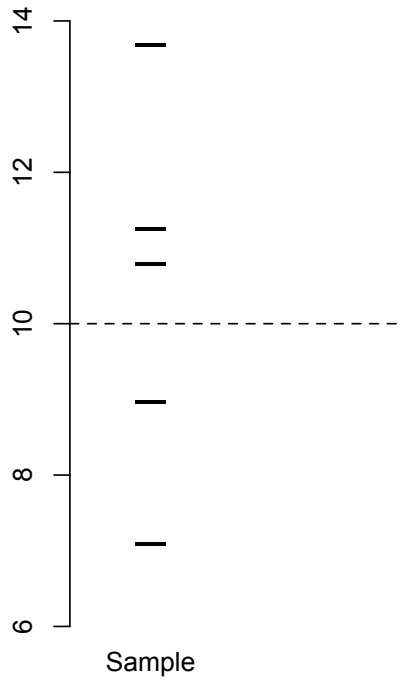
- Population distribution:
  Normal($\mu = 10, \sigma^2 = 4$).
- True population mean parameter: $\mu = 10$



Validity of Confidence Intervals

Example

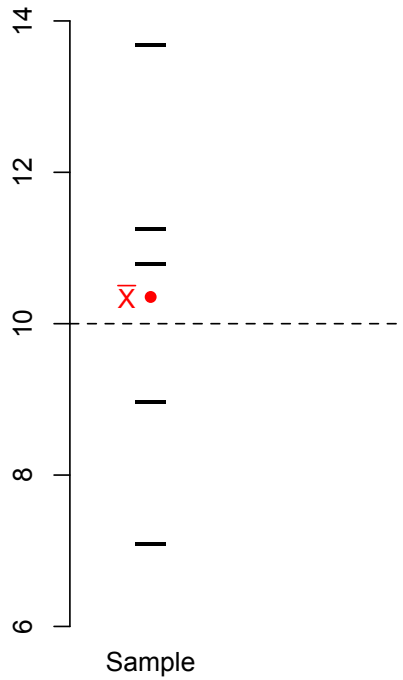
- Population distribution:
  Normal($\mu = 10, \sigma^2 = 4$).
- True population mean parameter: $\mu = 10$
- Draw a sample of $n = 5$ observations



Validity of Confidence Intervals

Example

- Population distribution: Normal($\mu = 10, \sigma^2 = 4$).
- True population mean parameter: $\mu = 10$
- Draw a sample of $n = 5$ observations
- Calculate sample mean $\overline{X}$ and sample variance s^2



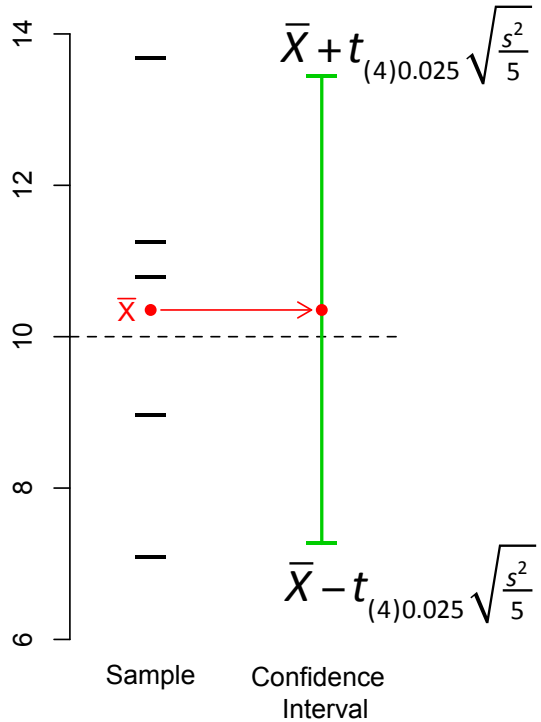
Validity of Confidence Intervals

Example

- Population distribution: Normal$(\mu = 10, \sigma^2 = 4)$.
- True population mean parameter: $\mu = 10$
- Draw a sample of $n = 5$ observations
- Calculate sample mean $\overline{X}$ and sample variance s^2
- Construct 95% confidence interval:

$$\left(\overline{X} - t_{(n-1)\alpha/2} \sqrt{\frac{s^2}{n}}, \overline{X} + t_{(n-1)\alpha/2} \sqrt{\frac{s^2}{n}} \right)$$

where $t_{(n-1)\alpha/2}$ is the critical value from a t-distribution with $(n-1)$ degrees of freedom.



Validity of Confidence Intervals

Example

- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)

Validity of Confidence Intervals

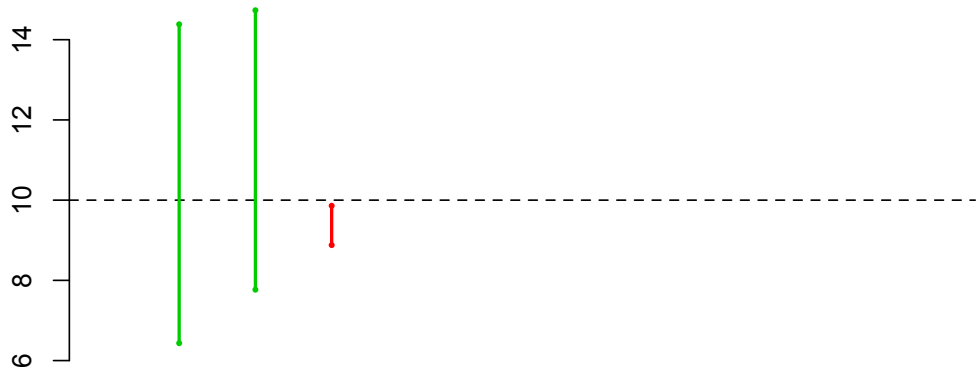
Example

- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)

Validity of Confidence Intervals

Example

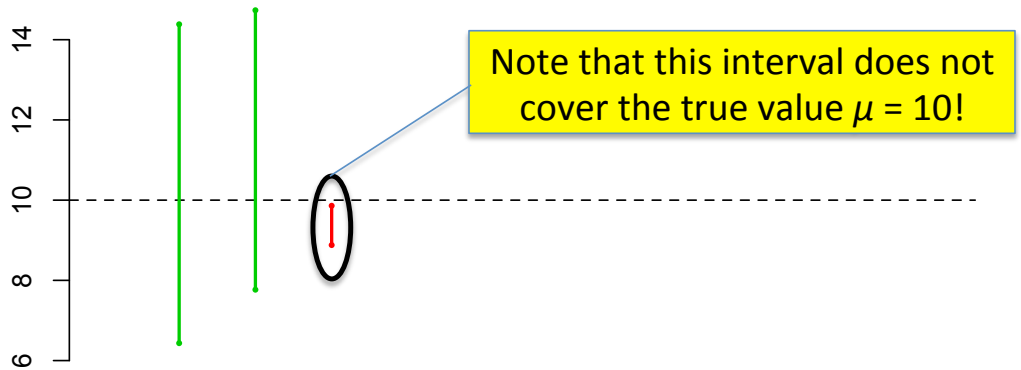
- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

Example

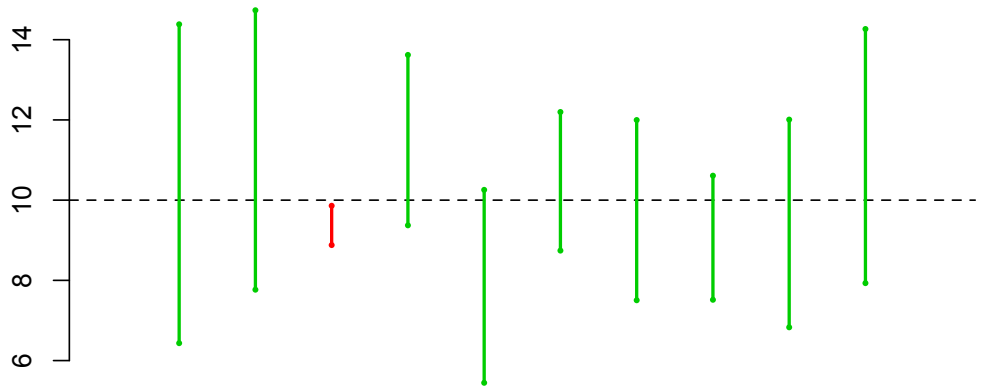
- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

Example

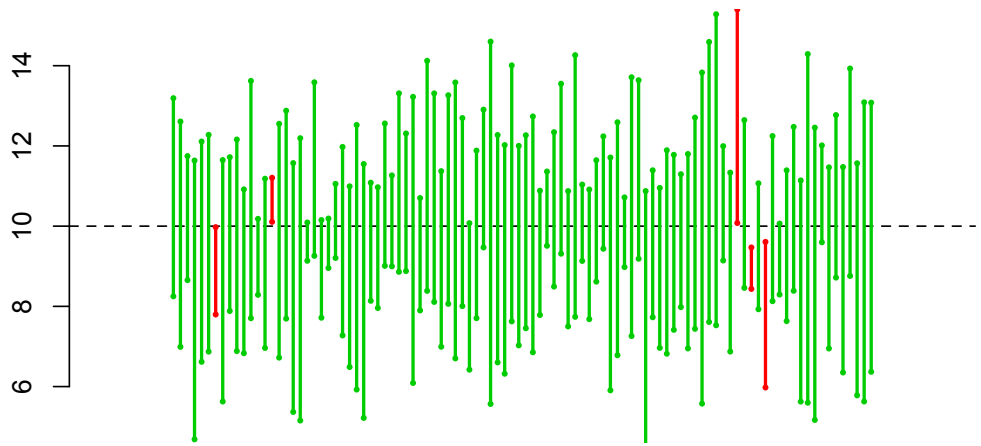
- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

Example

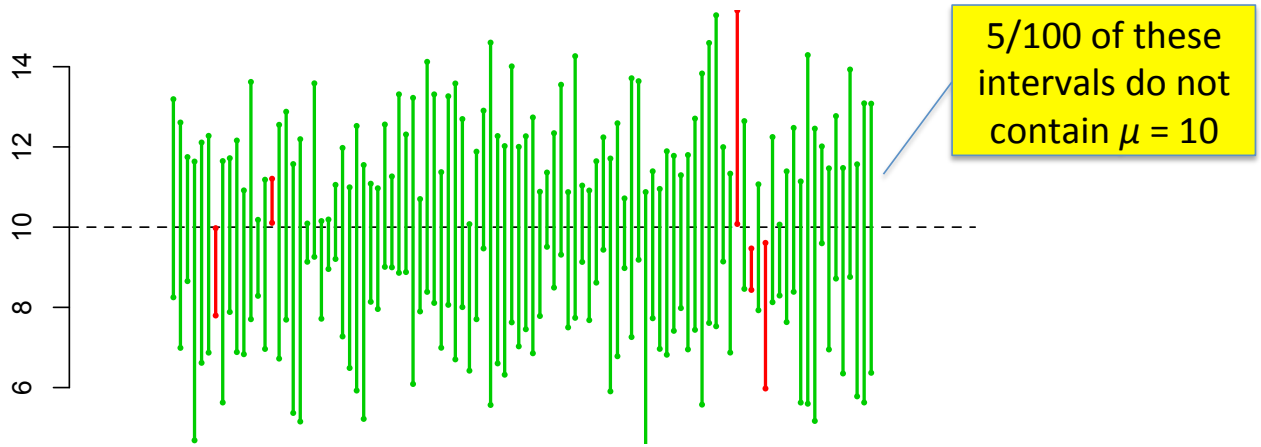
- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

Example

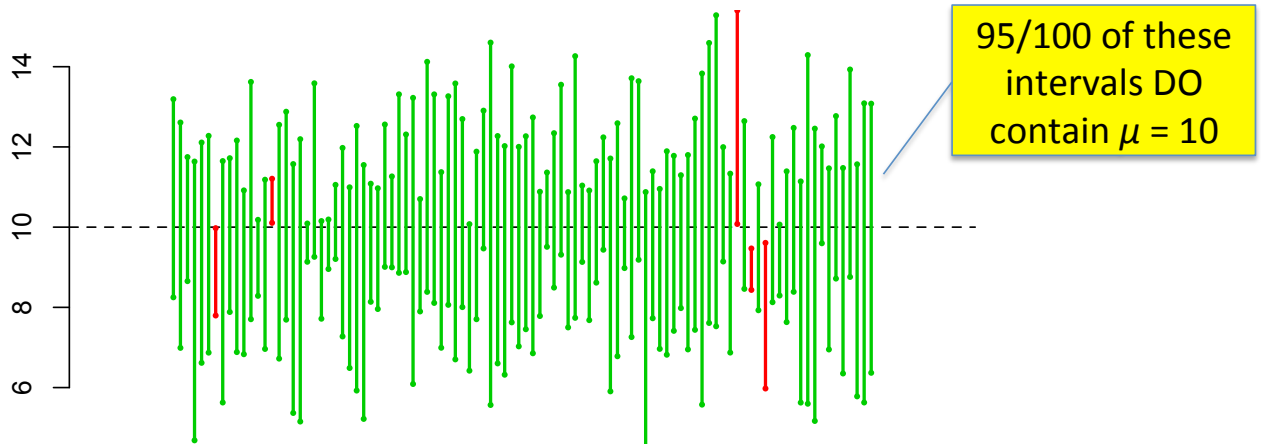
- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

Example

- Repeat this process many times
  - Draw new sample, construct confidence interval
- Examine how many of the resulting intervals contain the true value of $\mu = 10$
  - (We know the true value of μ because we simulated the data.)



Validity of Confidence Intervals

We can estimate coverage probability for a confidence interval method in a particular setting by

- Simulating a *really* large number of samples, each of size n (for instance, 10,000 samples)
- Constructing the confidence interval based on each sample
- Finding the proportion of resulting intervals that contain the true population parameter value

Validity of p-values

Validity of p-values:

A procedure for obtaining p-values is **valid** if the resulting p-value actually reflects the probability of obtaining results at least as extreme as the observed outcome when the null hypothesis is true.

To assess validity of p-values, we can:

- Use known mathematical theory...or
- Simulate data with a known true parameter value, and examine how frequently the resulting p-value (for testing the true null hypothesis) is less than some fixed probability u.
    - If the p-value method is valid, a proportion u of the p-values should be less than u when the null hypothesis is true.

Validity of p-values

Example

Suppose we find that when some particular null hypothesis is true, 10% of the time that we draw a sample and compute a p-value the resulting p-value is less than 0.05.

- In other words, 10% of the time, our p-value is saying 'we should only see a result at least this extreme 5% of the time'.

Validity of p-values

Example

Suppose we find that when some particular null hypothesis is true, 10% of the time that we draw a sample and compute a p-value the resulting p-value is less than 0.05.

- In other words, 10% of the time, our p-value is saying 'we should only see a result at least this extreme 5% of the time'.
- If our p-values were valid, exactly 5% of the time we should get a result *more extreme than* (resulting in a smaller p-value than) a p-value of 0.05.

Validity of p-values

Example

Suppose we find that when some particular null hypothesis is true, 10% of the time that we draw a sample and compute a p-value the resulting p-value is less than 0.05.

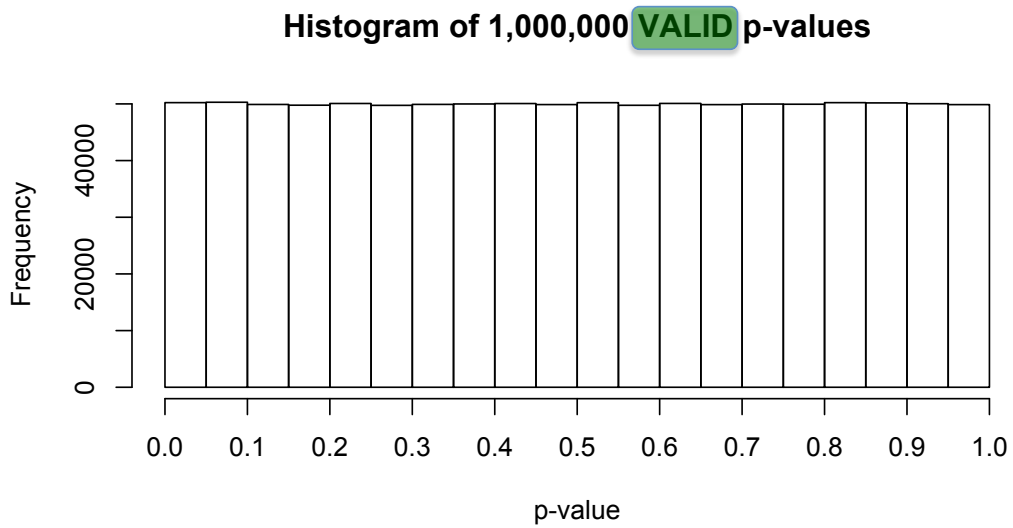
- In other words, 10% of the time, our p-value is saying 'we should only see a result at least this extreme 5% of the time'.
- If our p-values were valid, exactly 5% of the time we should get a result *more extreme than* (resulting in a smaller p-value than) a p-value of 0.05.

We can therefore conclude that this method of producing p-values is *not valid in this setting*.

Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

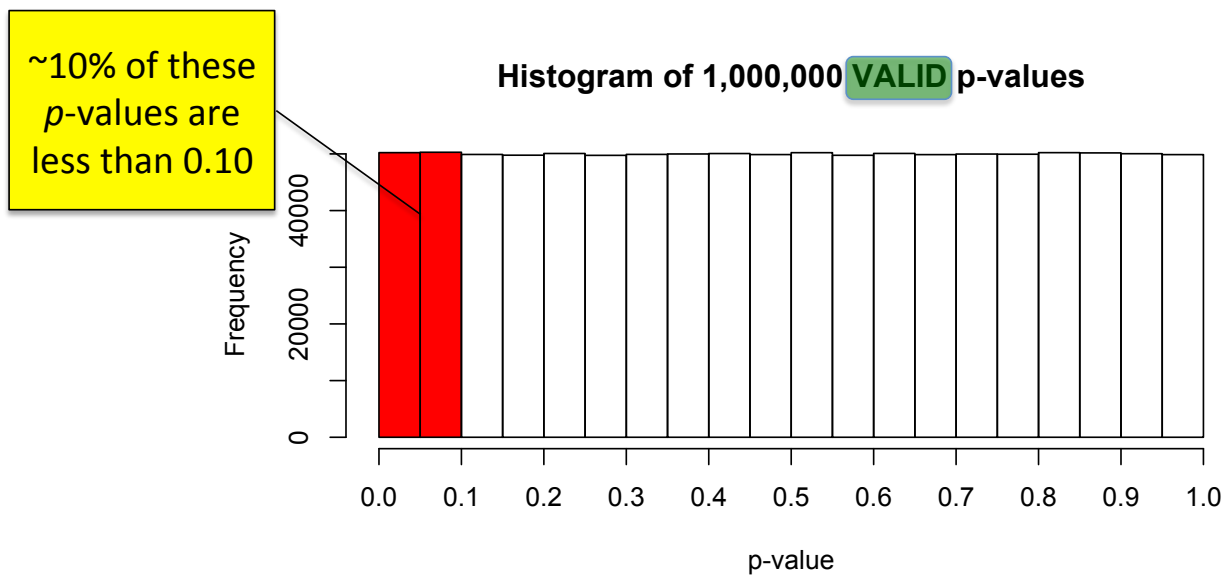
If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

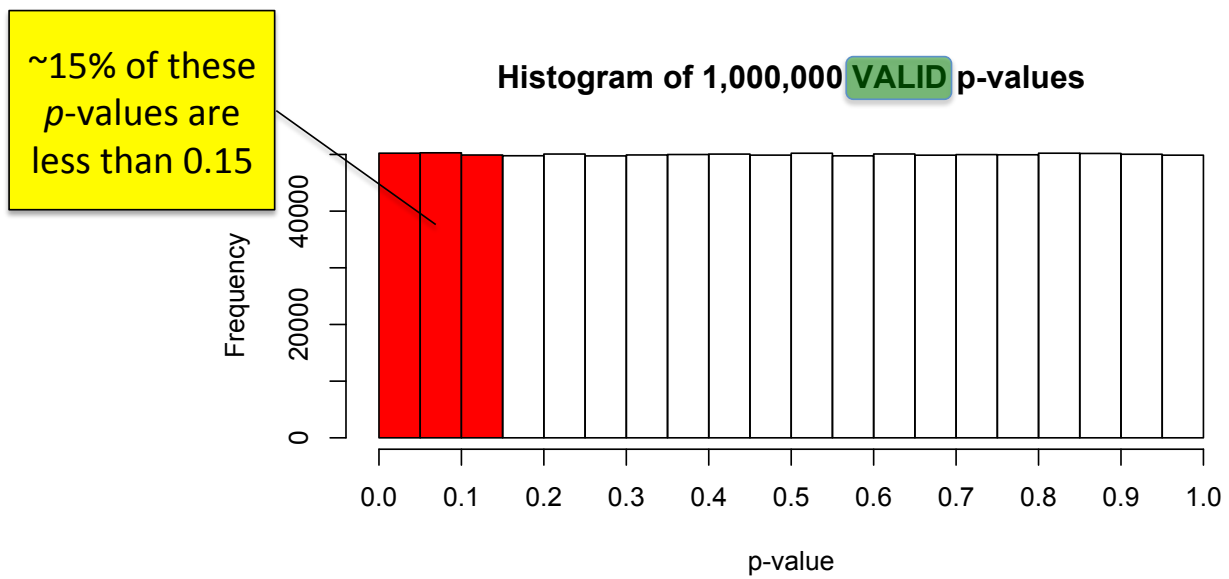
If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

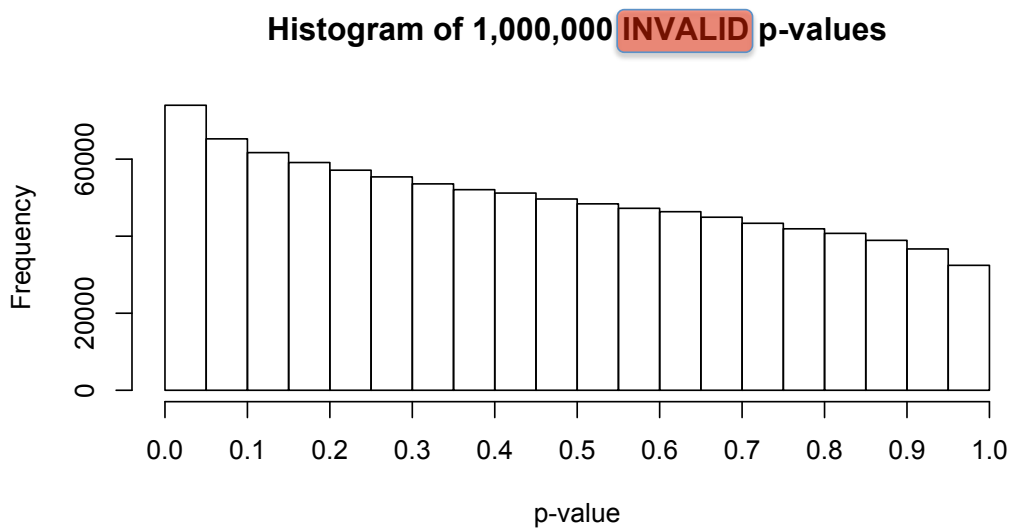
If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

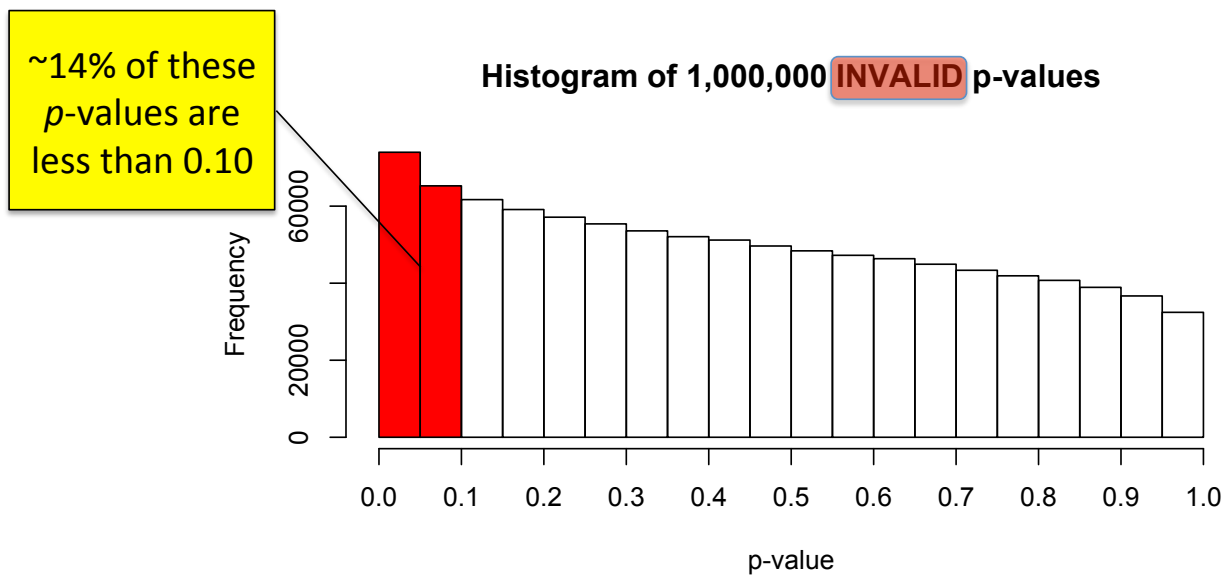
If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

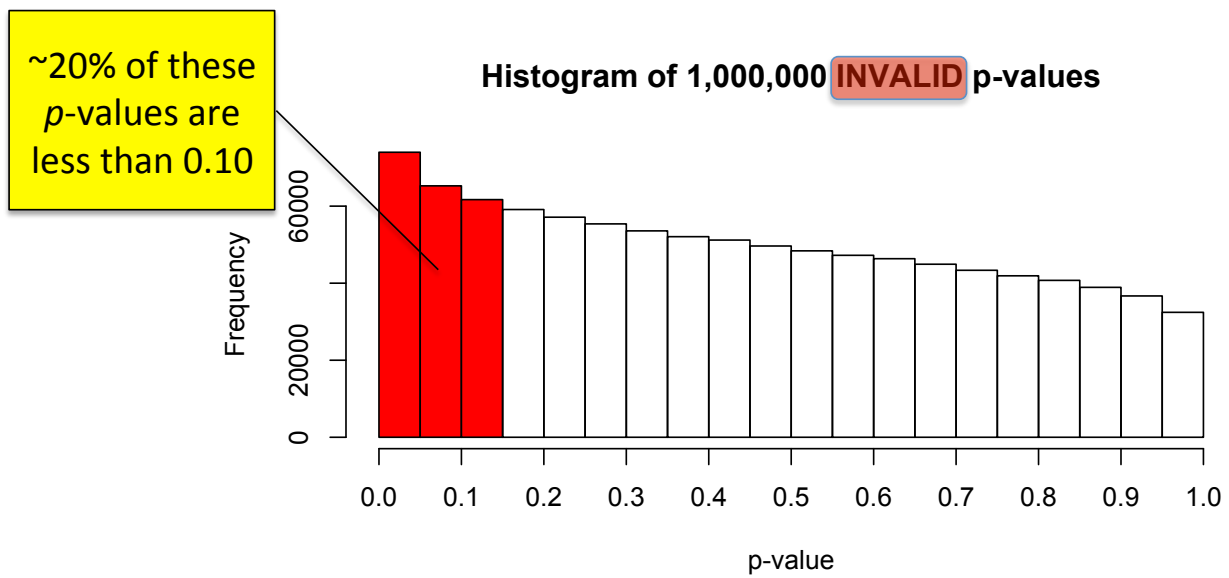
If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of p-values

If a method of calculating p-values is valid, a proportion u of the resulting p-values should be less than u when the null hypothesis is true.

If we generate a very large number of samples, and compute the p-value for testing the *true null hypothesis* for each sample, the histogram of all the resulting p-values should be uniform between 0 and 1.



Validity of Inference

Validity of confidence intervals, p-values, and hypothesis tests are all dependent on the same thing:

Validity of the reference null distribution

The **reference null distribution** (the distribution to which we compare our test statistic) is valid if the test statistic actually does have that distribution when the null hypothesis is true.

Validity of Inference

For example, when we perform a one-sample t-test, we compute a statistic

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{s^2/n}}$$

and then compare that statistic to a t-distribution with $(n-1)$ degrees of freedom (this is our reference null distribution for this statistic).

Does this statistic actually have a t-distribution with $(n-1)$ degrees of freedom when μ_0 is the true population mean (that is, when the null hypothesis $H_0 : \mu = \mu_0$ is true)?

- ***Yes***, if the population distribution is Normal and the observations in the sample are independent (and representatively sampled).
- ***Approximately***, no matter what the population distribution is, if the sample size is large and the observations in the sample are independent (and representatively sampled).
- ***No***, if the observations are not independent.