STATS
[dot] SEANDOLINAR [dot][com]

STATS

ONE MEAN Z-TEST [WITH R CODE]

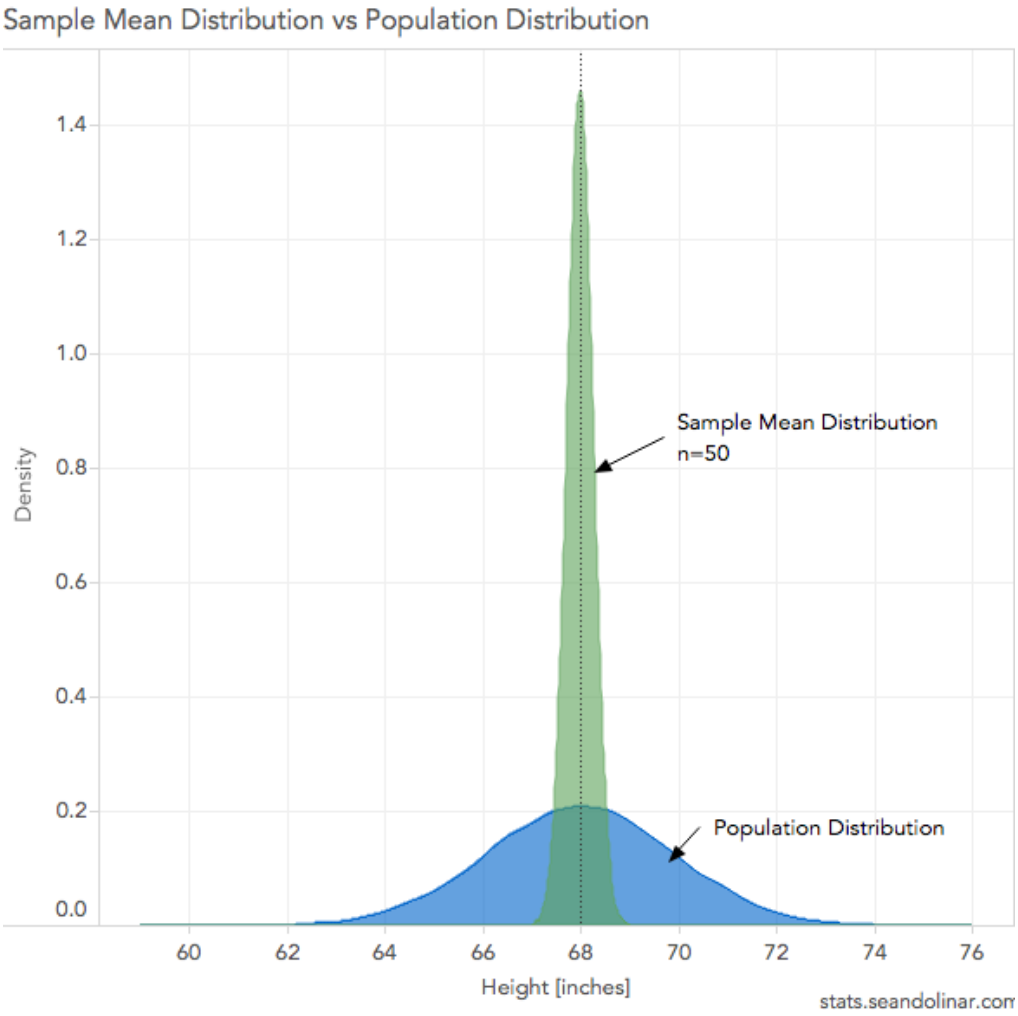
DECEMBER 19, 2014 | SEAN DOLINAR

I've included the full R code and the data set can be found on UCLA's Stats Wiki

Building on finding z-scores for individual measurement or values within a population, a z-test can determine if there is a statistically significance different between a sample mean and a population mean with a known population standard deviation. [Those conditions are essential for using this test.] The z-test uses z-scores and a normal distribution to determine the probability the sample mean is drawn randomly from a known population. If the test fails, the conclusion is that random sampling is likely to have produced this. If the test rejects the null hypothesis, then the sample is likely to be a result of non-random sampling [ie. like team captains picking the tallest kids for a basketball game in gym class].

The z-test relies critically on the central limit theorem, which basically states that if you take a n >= 30 sample a population [with any distribution] many times over, you'll get a normal distribution of the sample means. [This needs it's own post to explain fully, and there are interesting ways you can program R to illustrate this.] The sample mean distribution chart is shown below compared to the population distribution. The important concepts to notice here are:

- the area of both distributions is equal to 1
- the sample mean distribution is a normal distribution
- the sample mean distribution is tighter and taller than the population distribution

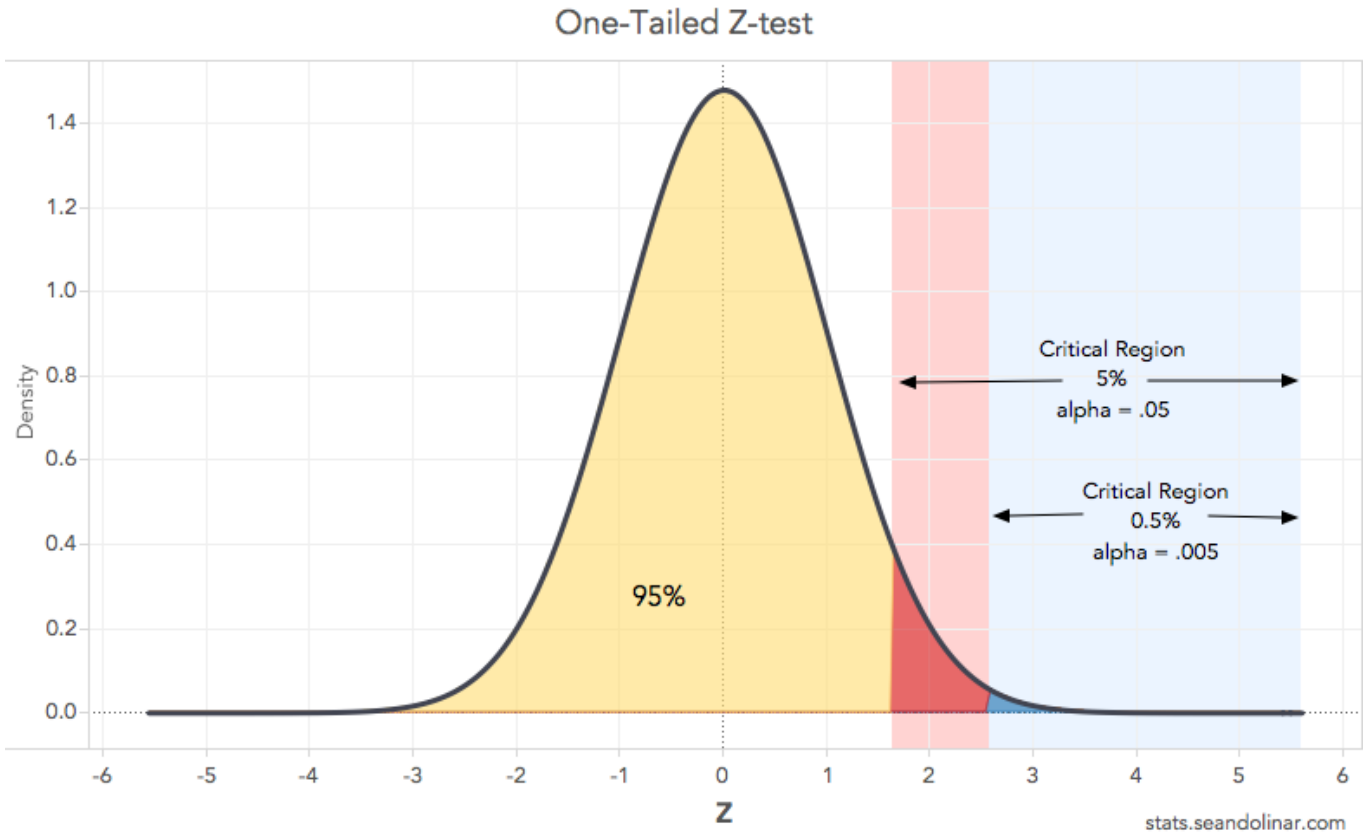


For the rest of this post, the sample mean distribution will be used for the z-test and it is also represent in green opposed to blue. Also the data I use in this post is height data from this data set. It represents the heights of 25,000 children from Hong Kong. The data doesn't reflect US adults, but it's a great normally distributed data set.

The goal of the z-test will be to test to see if a sample and its mean are randomly sampled from the population or if there's some significant difference. For example, you could use this test to see if the average height of NBA players is statistically significantly different than the general population. While the NBA example is pretty common sense, not every problem will be that clear. Sample size [like in many hypothesis tests] is a huge factor. Small sample sizes require huge differences between the sample mean and the population mean to be significant.

For a one-mean z-test, we will be using a one-tail hypothesis test. The null hypothesis will be that there is NO difference between the sample mean and the population mean. The alternate hypothesis will test to see if the sample mean is greater. The null and alternate hypotheses are written out as:

- $latex H_0: \bar{x} = \mu&s=2$
- $latex H_A: \bar{x} > \mu&s=2$



The graph above shows the critical regions for a right-tailed z-test. The critical regions reflect areas where the z-stat has to fall in order for the test to reject the null hypothesis. The critical regions are defined because they represent a probability less the the stated confidence level. For example the critical region for 95% confidence level only has an area [probability] of 5%. If the sample mean is the same as the population mean, there's a 5% chance it was drawn by random chance. This concept is the basis for almost every hypothesis test.

The z-test uses the z-stat, which is calculated analogously to the z-score the difference being it uses standard error instead of standard deviation. These two concepts are similar; The standard deviation applies to the 'spread' of the blue population distribution, while the standard error applies to the 'spread' of the green sample mean distribution. The z-stat is calculated as:

$$latex z = \frac{\bar{x} – \mu}{\sigma/\sqrt{n}} &s=2$$

The higher the z-stat is the more certainty there is that the sample mean and the population are different. There are three things make the z-stat larger:

- a bigger difference between sample mean and population mean
- a small population standard deviation
- a larger sample size

Example

I have two sets of sample from the data set: one is entirely random and the other I weighted heavily towards taller people. The null hypothesis would be that both there's no difference between the sample mean and the population mean. The alternate would be that the sample mean is greater than the population mean. The weighted sample would be the sample if you

were evaluating the mean height of a basketball team vs the general population. Here are the two sets of an n=50 sample and R code on how I constructed them using a set random seed of 123.

Unbiased random sample

69.895	65.618	65.519	69.998	68.668	64.646	66.320	68.308	68.306	65.654
70.424	68.833	71.467	69.196	66.804	64.168	69.177	67.772	67.473	68.805
69.075	64.989	67.670	71.100	67.889	67.135	69.106	71.028	70.869	64.689
65.052	68.297	72.325	66.756	66.340	62.508	70.123	65.961	67.632	67.973
69.914	68.907	67.275	69.646	68.584	70.003	69.201	69.028	70.492	68.311

Tall-biased random sample

69.660	67.413	69.098	66.769	66.196	71.573	68.603	66.693	68.510	68.898
65.949	68.912	67.967	68.424	70.869	66.625	69.872	71.642	69.467	65.984
66.723	67.901	68.124	64.620	68.055	67.826	68.541	68.333	69.652	70.508
65.826	66.603	67.912	67.373	72.062	68.811	67.571	70.042	69.512	69.955
70.545	69.075	69.080	69.273	70.470	70.572	69.946	68.859	69.765	66.972

```r
#unbiased random sample
set.seed(123)
n <- 50
height_sample <- sample(height, size=n)
sample_mean <- mean(height_sample)

#tall-biased sample
cut <- 1:25000
weights <- cut^.6
sorted_height <- sort(height)
set.seed(123)
height_sample_biased <- sample(sorted_height, size=n, prob=weights)
sample_mean_biased <- mean(height_sample_biased)
```

The population mean is 67.993, the first unbiased sample is 68.099, and the tall-biased group is 68.593. Both samples are higher than the than the population mean, but are both significantly higher than the mean? To figure this out, we need to calculate the z-stats and find out if those z-stats fall in the critical region using the equation:

$latex z = \frac{\bar{x} – \mu}{\sigma/\sqrt{n}} &s=2$

We can substitute and calculate with the population standard deviation [σ] = 1.902:

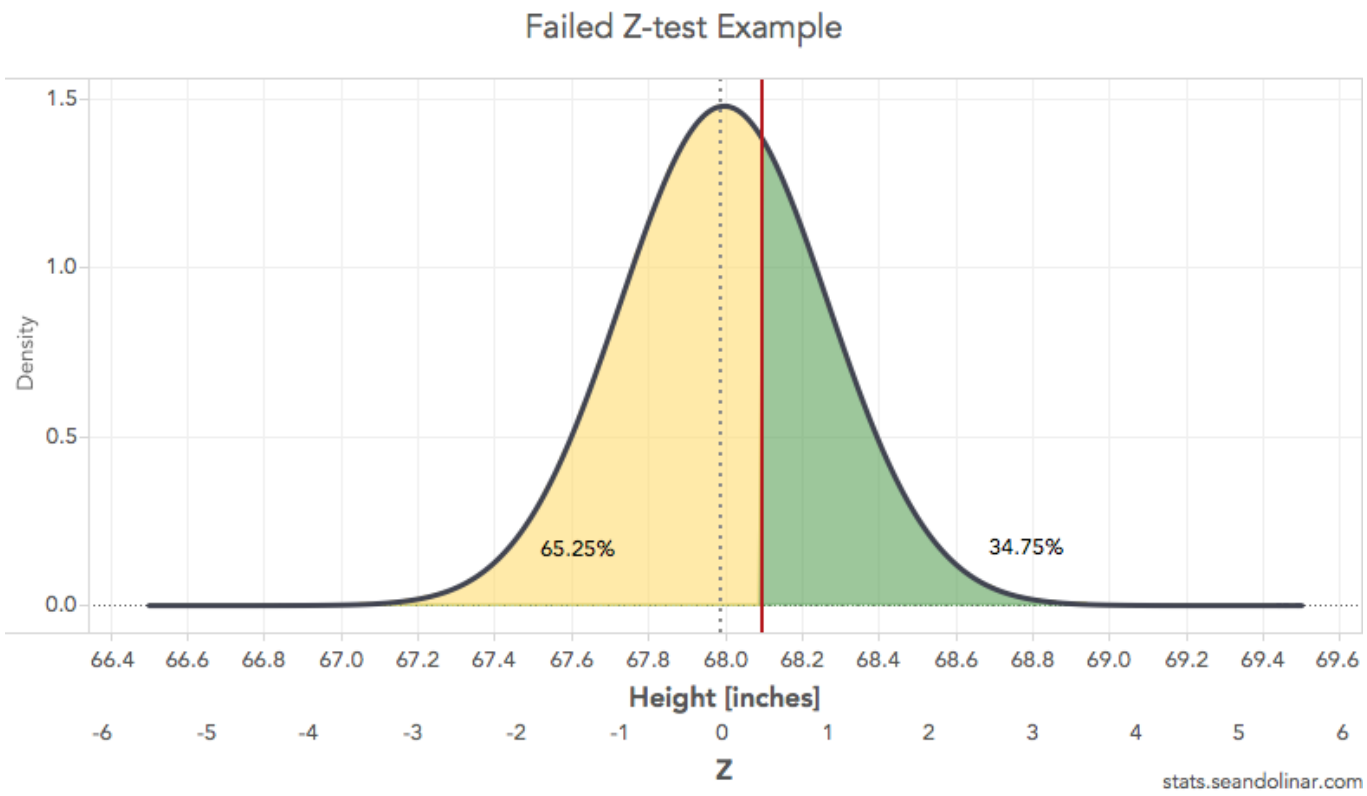
$latex z_{unbiased} = \frac{68.593 – 67.993}{1.902/\sqrt{50}} = 0.3922 \ \ \ z_{tall-biased} = \frac{68.099 – 67.993}{1.902/\sqrt{50}} = 2.229 &s=0$

```r
#random unbiased sample
#z-stat calculation
sample_mean
z <- (sample_mean - pop_mean)/(pop_sd/sqrt(n))

#tall-biased sample
z <- (sample_mean_biased - pop_mean)/(pop_sd/sqrt(n))
```

Quickly, knowing that the critical value for a one-tail z-test at 95% confidence is 1.645, we can determine the unbiased random sample is not significantly different, but the tall-biased sample is significantly different. This is because the z-stat for

the unbiased sample is less than the critical value, while the tall-biased is higher than the critical value.



Plotting the z-test for the unbiased sample, the area [probability] to the right of the z-stat is much higher than the accepted 5%. The larger the green area is the more likely the difference between the sample mean and the population mean were obtained by random chance. To get a z-test to be significant, you want to get the z-stat high so that the area [probability] is low. [In practice, this can be done by increasing sample size.]

The tall-baised sample mean's z-stat creates a plot with much less area to the right of the z-stat, so these results were much less likely to be obtained by chance. The p-values can be obtained by calculating the area to right of the z-stat. The R code below summarizes how to do that using R's 'pnorm' function.

```
#calculating the p-value
p_yellow2 <- pnorm(z)
p_green2 <- 1 - p_yellow2
p_green2
```

The p-value for the unbiased sample is .3474 or there's a 34.74% chance that the result was obtained due to random chance, while the tall-biased sample only have a p-value of .01291 or a 1.291% chance being a result of random chance. Since the p-value tall-biased sample is less than the .05, the null hypothesis is rejected, but the since the unbiased sample's p-value is well above .05, the null hypothesis is retained.

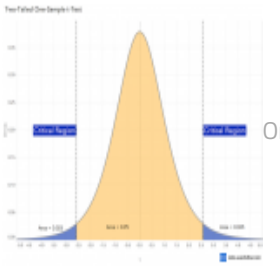
What the one-mean z-test accomplished was telling us that a simple random sample from a population wasn't really that different from population, while a sample that wasn't completely random but was much taller than the overall population was shown to be different. While this test isn't used often, the principles of distributions, calculating test stats, and p-values have many applications with in the statistics universe.

FEATURED HEIGHT MEAN NORMAL DISTRIBUTION STATS Z-STAT Z-TEST

PREVIOUS POST
Calculating Z-Scores [with R code]

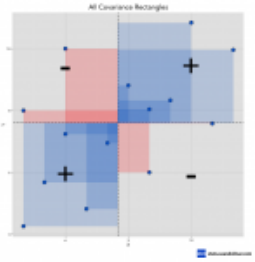
NEXT POST
2015 Steelers-Ravens Playoff Twitter Infographics

RELATED POSTS



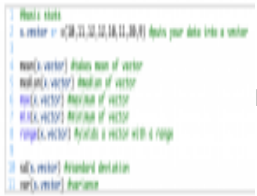
One-Sample t-Test [With R Code]

Open Graph Test



Covariance — Different Ways to Explain or Visualize It

Making a Covariance Matrix in R



R Bootcamp — A Quick Introduction

@seandolinar