Testing with Normal Distributions

ECE 341: Random Processes

Lecture #18

note: All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

Testing with Normal Distributions

If a variable has a normal distribution, you can determine certain probabilities. This lecture covers three of these:

- Single-sided confidence interval
- Two-sided confidence interval
- Comparison of Two Distributions
  - False Positives
  - False Negatives

For illustration purposes, consider the gain of 62 Zetex1051a transistors:

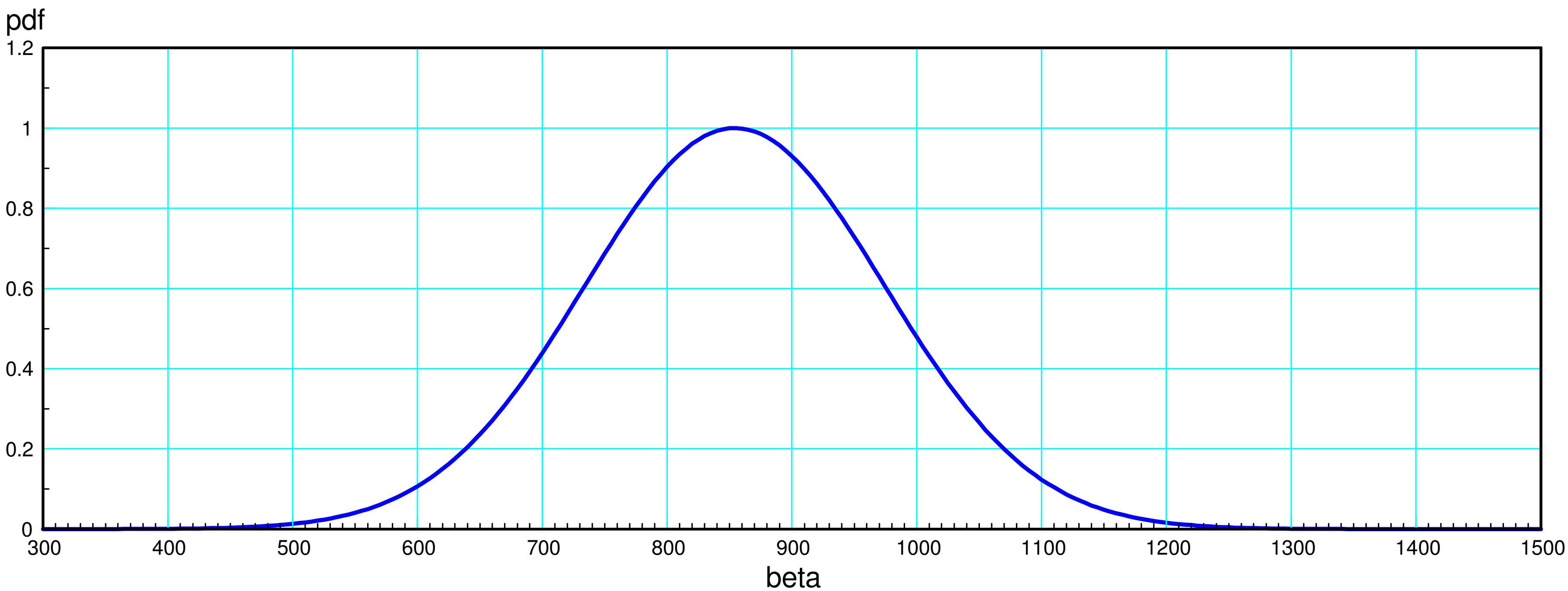
```
915, 602, 963, 839, 815, 774, 881, 912, 720, 707, 800, 1050, 663, 1066, 1073,
802, 863, 845, 789, 964, 988, 781, 776, 869, 899, 1093, 1015, 751, 795, 776,
860, 990, 762, 975, 918, 1080 774, 932, 717, 1168, 912, 833, 697, 797, 818,
891, 725, 662, 718, 728, 835, 882, 783, 784, 737, 822, 918, 906, 1010, 819,
955, 762
```

From this data, the mean and standard deviation of these transistors can be found

```
x = mean(beta)          x =  854.1290

s = std(beta)           s =  120.2034
```

From the Central Limit Theorem, with a sample size of 62, these approach a normal distribution with the same mean and standard deviation.



Normalized pdf for the current gain, beta (hfe).

With this curve, we can answer several questions.

Single-Sided Test:

The data sheets state that the gain is at least 300.

- What is the probability that a transistor will have a gain less than 300?

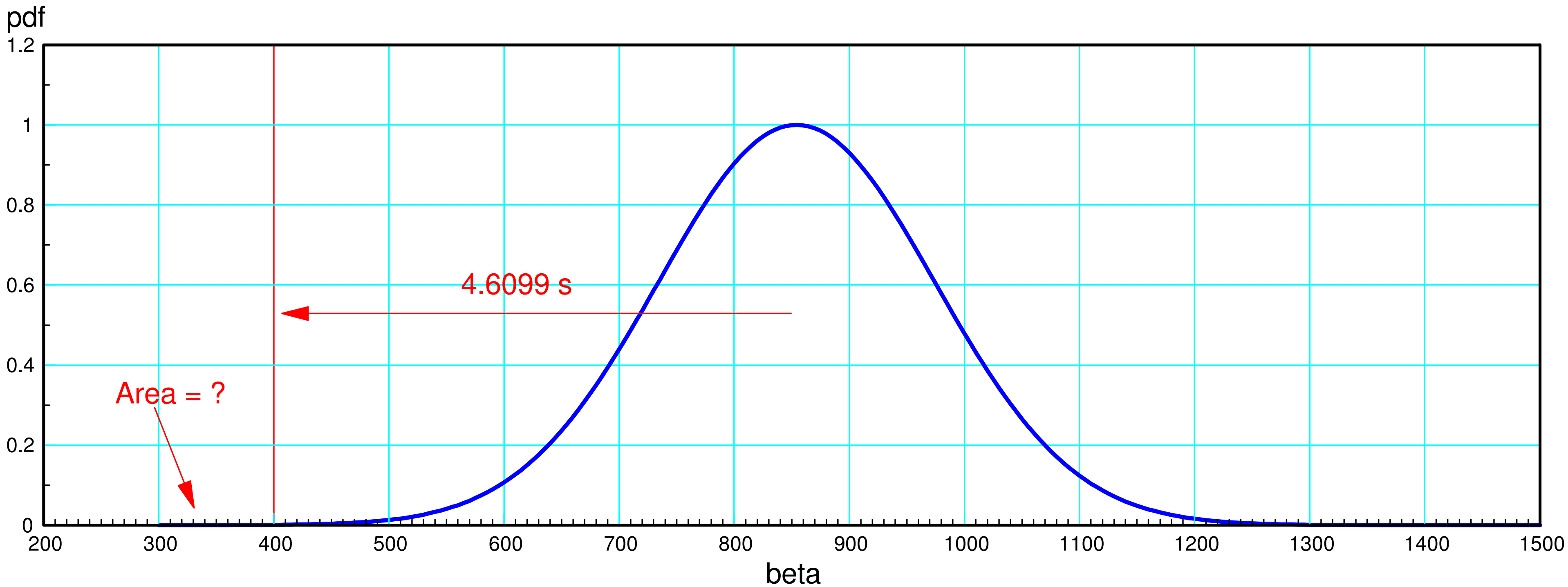
Find the z-score (distance to the mean in terms of standard deviations)

$$z = \left(\frac{\mu - 300}{\sigma}\right) = \left(\frac{854.129 - 300}{120.2013}\right) = 4.6099$$

Use a standard normal table to convert to a probability

- $p < 0.00003167$
- StatTrek also works

Deviations	+0	+1	+2	+3	+4	+5
Area of Tail	0.5	0.1587	0.0227	0.001349	3.167 10-5	2.866 10-7

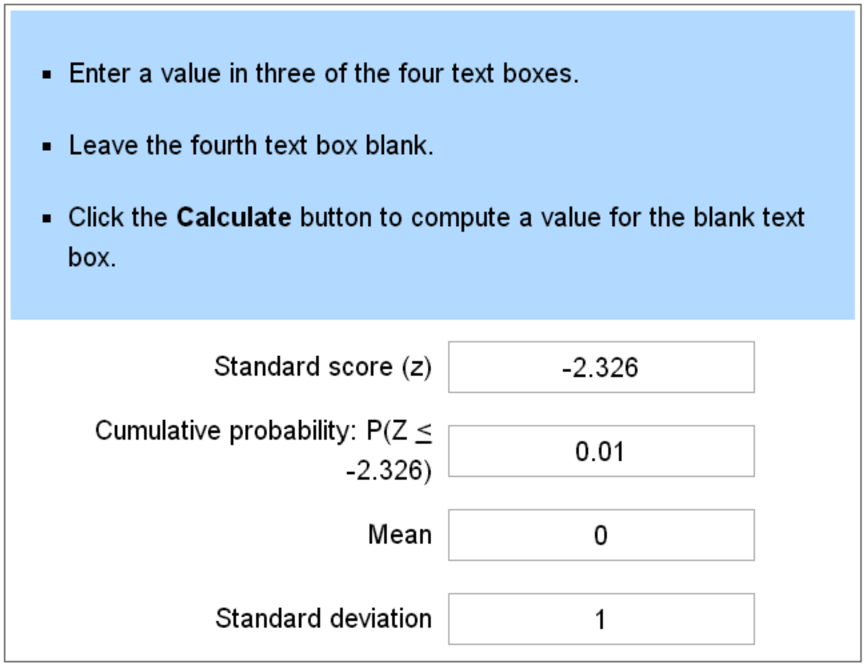


Probability that a transistor has a gain less than 300 is less than 0.00003167

Essentially, the manufacturer is cautions in the claims.

Example 2: Determine the gain that 99% of all transistor will meet or exceed.

Solution: Use StatTrek to determine how many standard deviations you have to go for the area to be 0.01 (1%)



- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

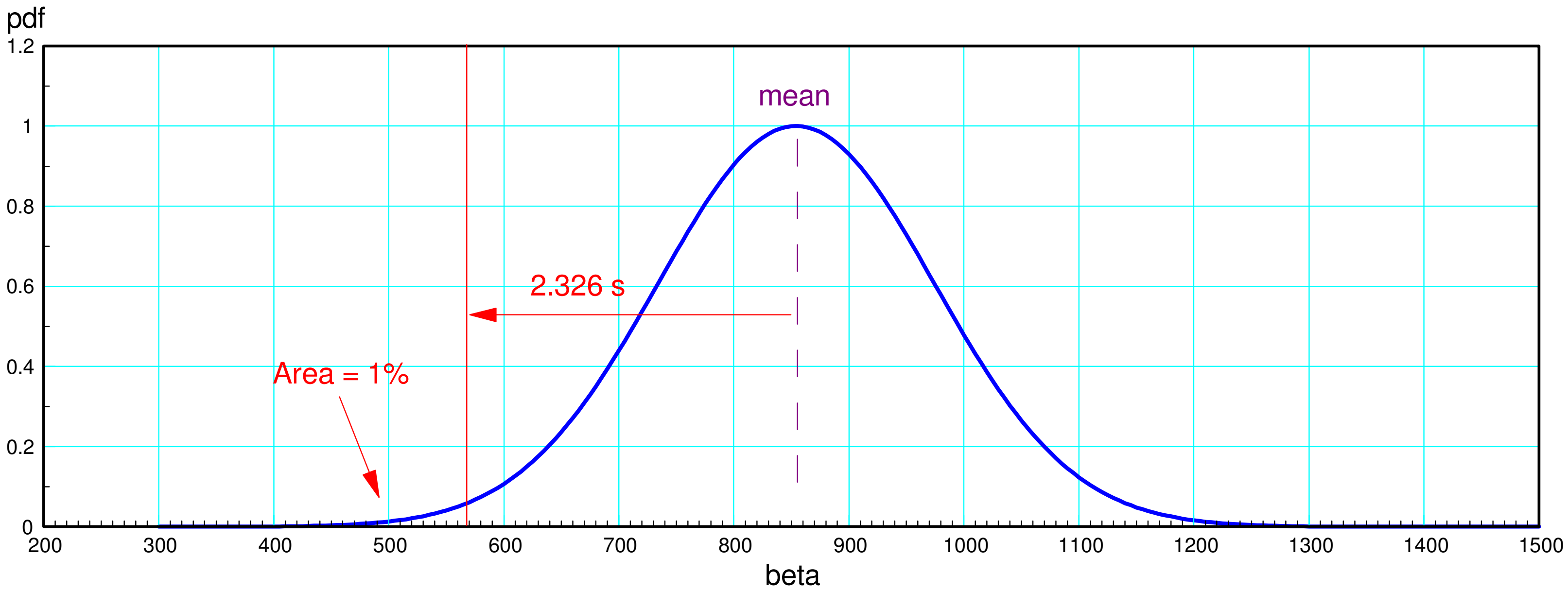
Standard score (z)	-2.326
Cumulative probability: P(Z ≤ -2.326)	0.01
Mean	0
Standard deviation	1

What this means is

- Go 2.326 standard deviations to the left of the mean
- The area under the curve (i.e. the probability) will be 1%

Translating to gain:

$$\beta > \mu - 2.326\sigma = 574.578 \qquad (\ p = 0.99\ )$$



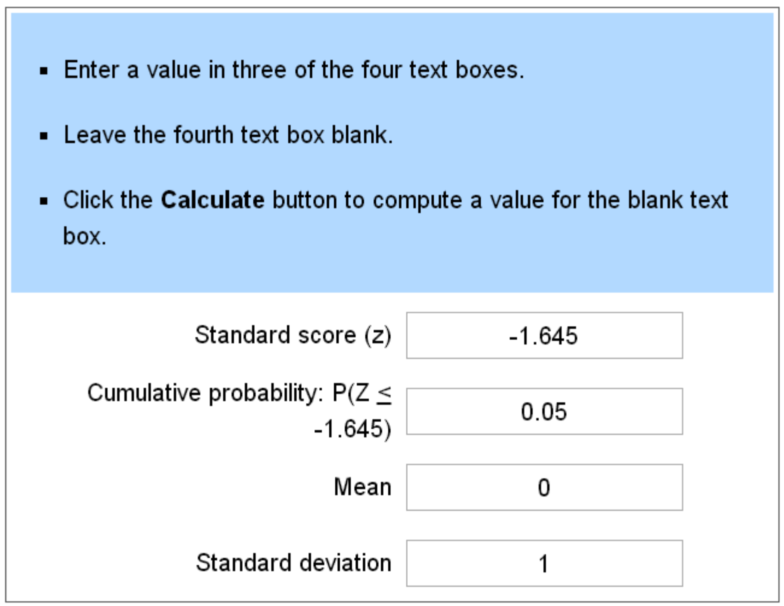
Single-Sided Test: Any given transistor will have a gain > 574 with a probability of 0.99

Two-Sided Tests (Confidence Intervals)

Example 3: Determine the 90% confidence interval for the gain of a given transistor.

Solution:

- Area in the middle = 90%
- Each tail = 5% (two tails)
- Find the z-score for 5% tails

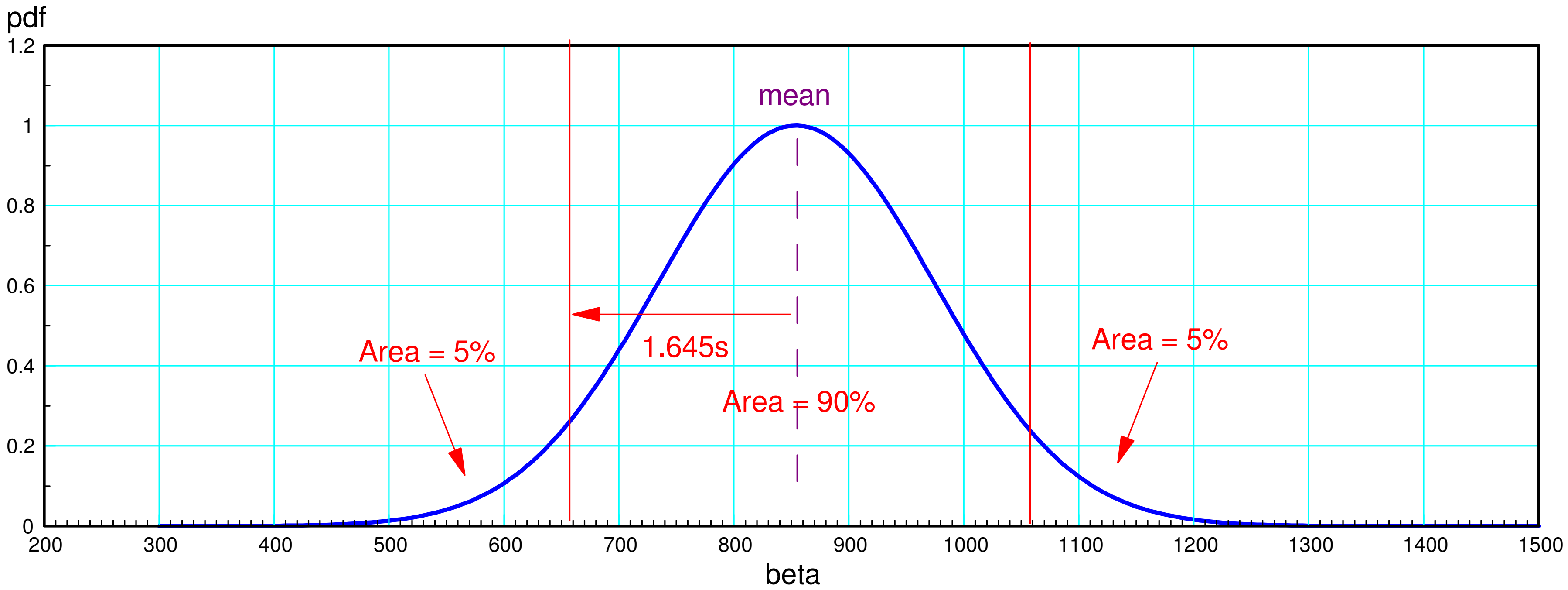


The 90% confidence interval is then

$$\mu - 1.645\sigma < \beta < \mu + 1.645\sigma \qquad p = 0.9$$

Plugging in numbers

$$656.39 < \beta < 1051.9 \qquad p = 0.9$$



90% Confidence Interval for Transistor Gain: Note that each tail has an area of 5%

Example 4: What is the 99% confidence interval?

- Repeat but with each tail being 0.5%:

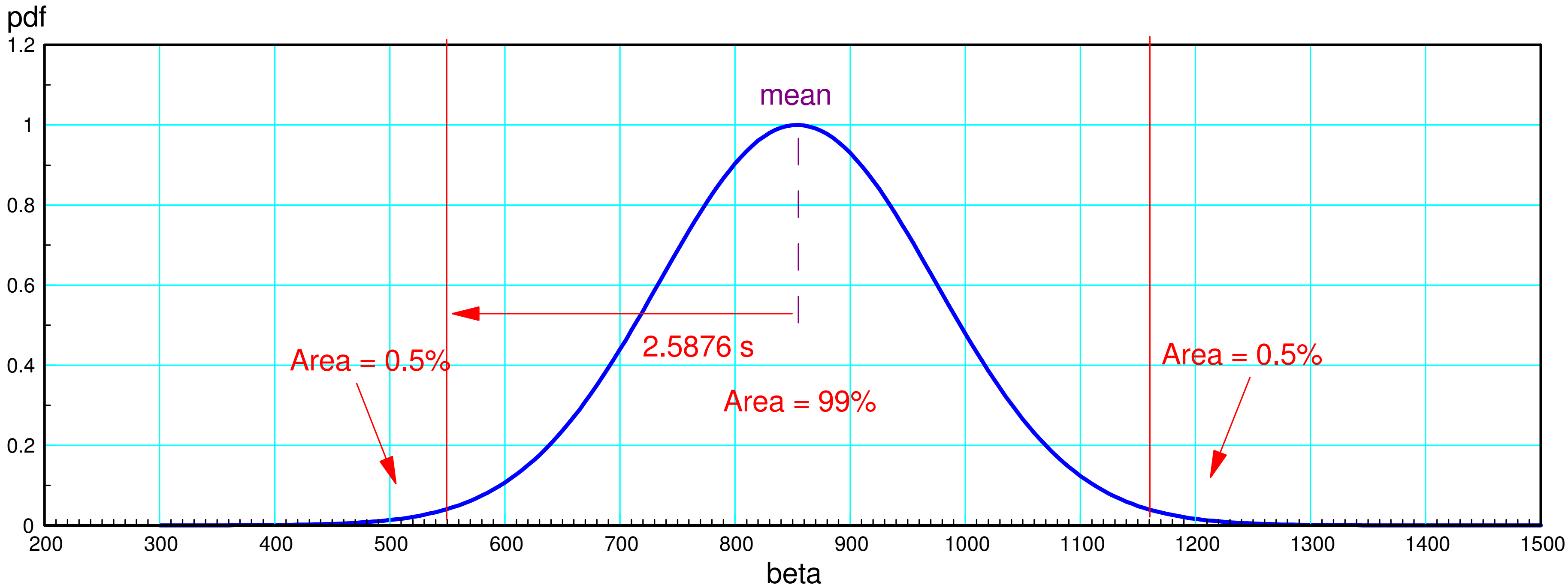
  - Enter a value in three of the four text boxes.
  - Leave the fourth text box blank.
  - Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	-2.576
Cumulative probability: P(Z ≤ -2.576)	0.005
Mean	0
Standard deviation	1

The 99% confidence interval is then

$$\bar{x} - 2.576s < \beta < \bar{x} + 2.576s$$

$$544.84 < \beta < 1163.8$$



99% confidence integral for the current gain

Example 5: What is the 100% confidence interval?

- Repeat but with each tail being 0%
- The z-score is infinity

$$-\infty < \beta < \infty \qquad\qquad p = 100\%$$

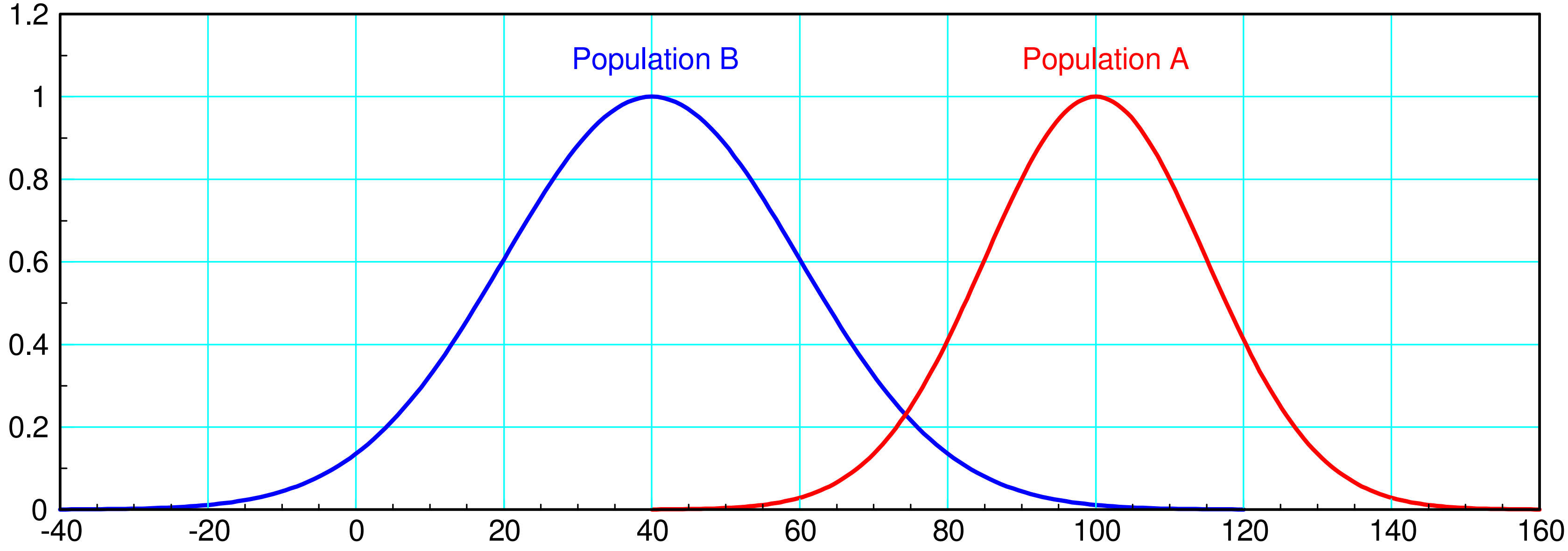
100% probability makes no sense.

- Nothing is 100% certain
- We aren't even 100% certain the world exists

Comparing Two Populations

- Let A have a normal distribution with mean μ_a and variance σ_a^2
- Let B have a normal distribution with mean μ_b and variance σ_b^2
- Let (a, b) be random samples from population A and B

What is the chance that a > b?

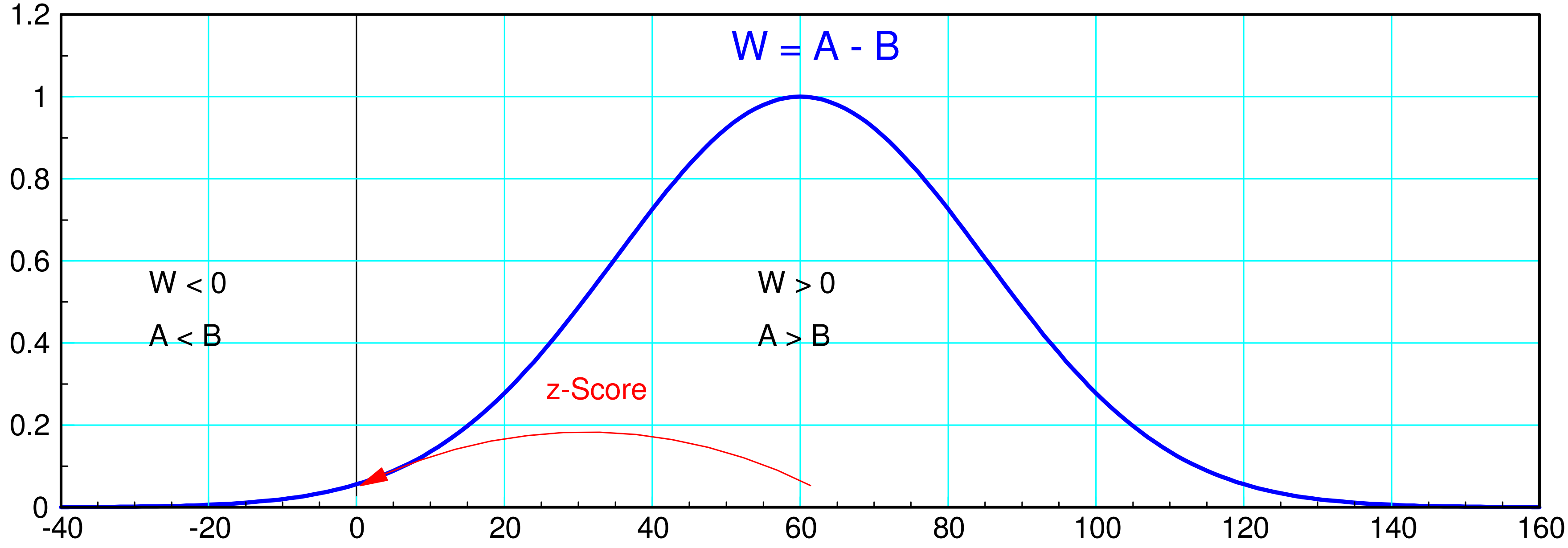


Solution:

Create a new variable, W = A - B

- W has a normal distribution
- $\mu_w = \mu_a - \mu_b$
- $\sigma_w^2 = \sigma_a^2 + \sigma_b^2$

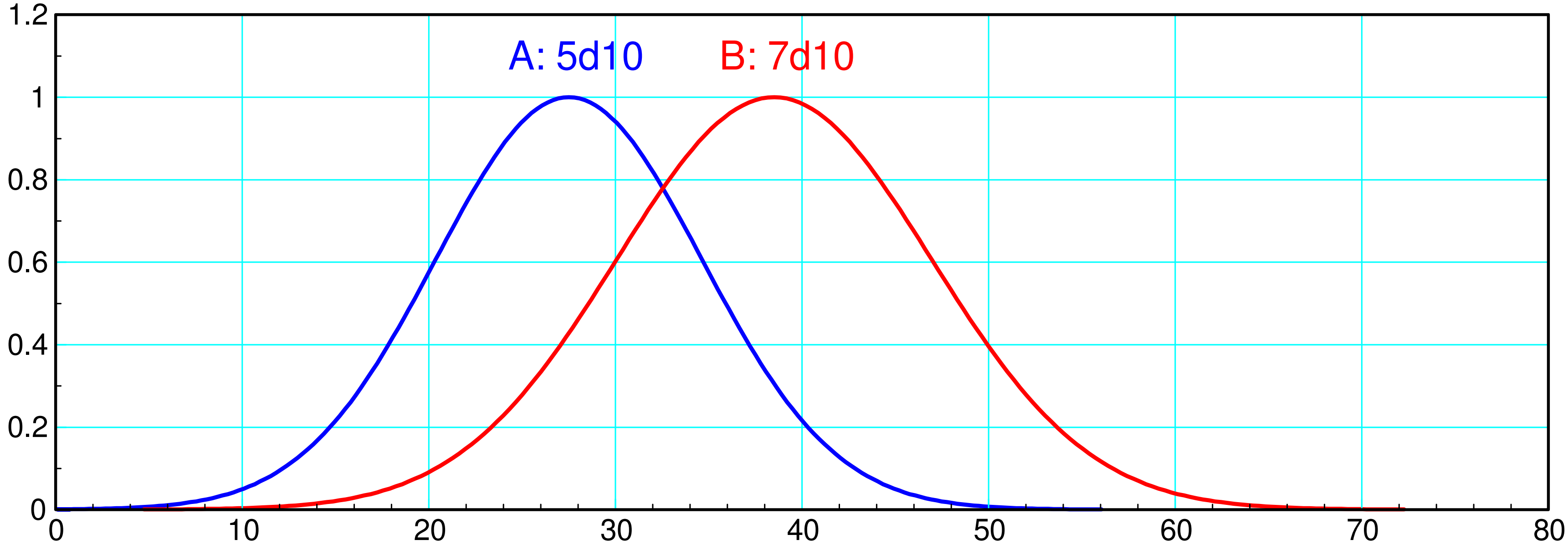
Find the area to the right of W=0



Example: 5d10 vs. 7d10

- Player A rolls five 10-sided dice
  - 5d10
  - Score = [5, 50]
- Player B rolls seven 10-sided dice
  - 7d10
  - Score = [7, 70]

What is the chance that A > B?



Solution (normal approximation)

Find the mean and variance of A and B

Find the mean and variance of W = A - B

	Mean	Variance
10-sided die (d10)	5.5	10.1852
A: 5d10	27.5	50.9259
B: 7d10	38.5	71.2963
W = A - B	-11.0	122.22

Find the area of W > 0

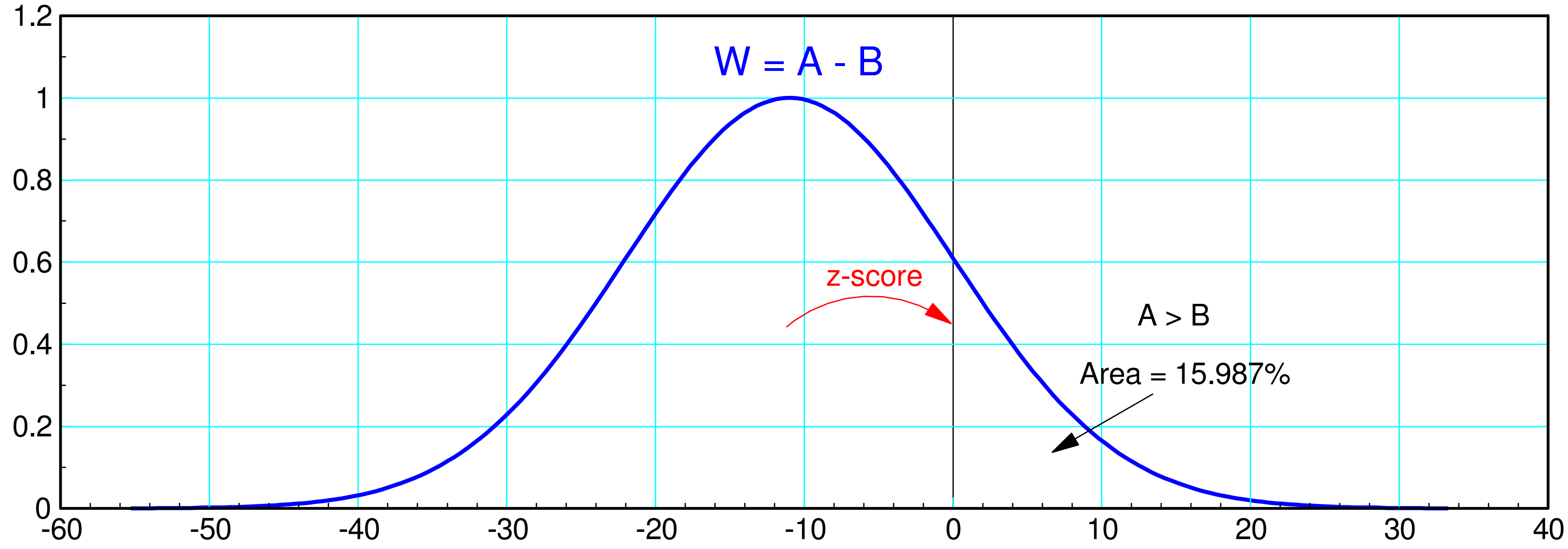
- A > B (ignores ties)

z-score

$$z = \left(\frac{\mu_w - 0}{\sigma_w}\right) = \left(\frac{-11}{\sqrt{122.22}}\right) = -0.9950$$

From a normal table (or StatTrek), the tail has an area of 0.15987

- A has a 15.987% chance of winning



Check: Monte-Carlo Simulation

```
W = 0;
T = 0;
L = 0;
for n=1:1e6
   A = sum(ceil(10*rand(1,5)));
   B = sum(ceil(10*rand(1,7)));
   if(A>B) W = W + 1;  end
   if(A==B) T = T + 1; end
   if(A<B) L = L + 1; end
end
disp([W,T,L])

Wins          Ties        Losses
125382        22259       852359

12.53%         2.22%      85.23%
```

Result is better with continuous distributions

- no chance of a tie

Result is better with more die rolls

- central limit theorem gets more accurate

Testing: (False Positives, False Negatives)

Assume

- A population falls into two groups, A and B.
- You run a test and the output of that test differs for each group:
  - Population A has its mean and variance
  - Population B has its mean and variance.

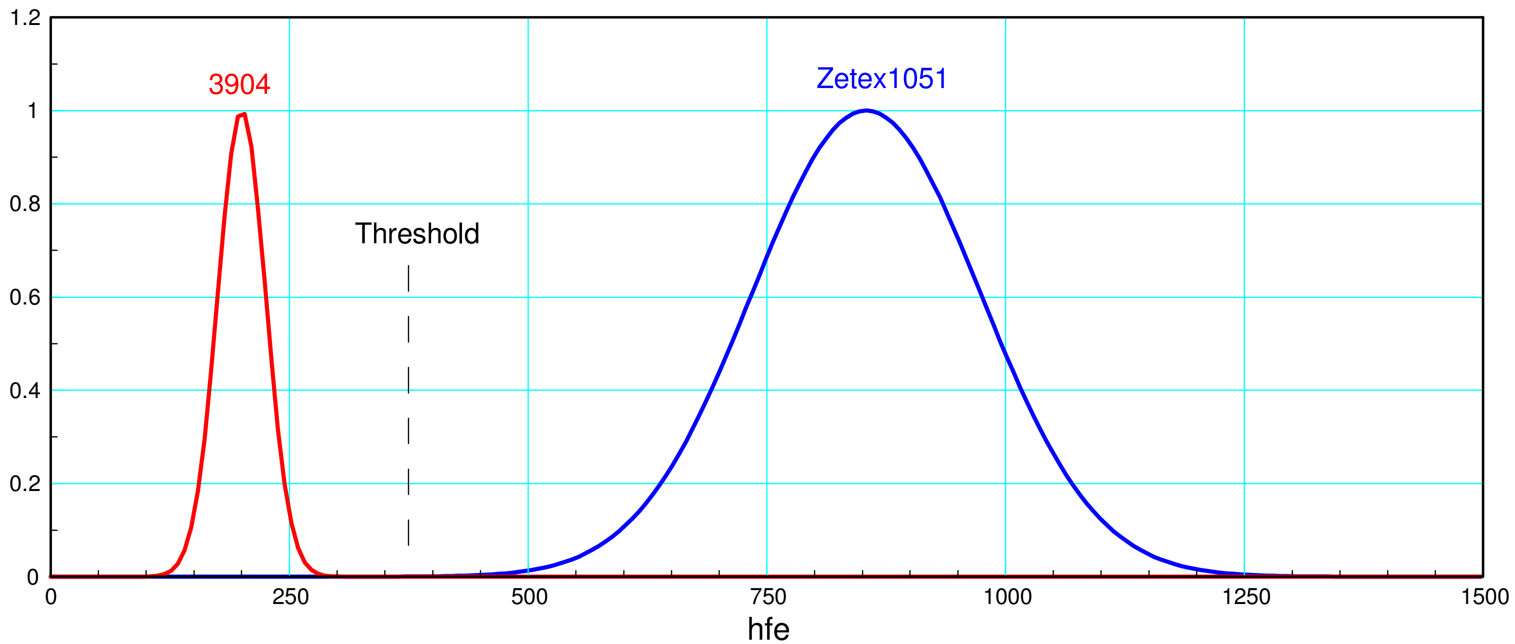
Given a test outcome, determine whether that person belongs to group A or B.

For example, let

- A be Zetex1051a transistors with a current gain
  - $\mu_A = 854.1290$ mean
  - $\sigma_A = 120.2034$ standard deviation
- B be a 3904 transistor with
  - $\mu_B = 200$ mean
  - $\sigma_B = 25$ standard deviation

Assume a bag has an assortment of both transistors without labels. Can you determine what type of transistor it is by measuring the gain?

If you graph the normalized pdf's (shown below), you can see that, yes, it is easy to tell which population a given transistor comes from



Normalized pdf of a 3904 transistor and Zetex 1051a transistor

Since the two pdf's are distinct, pick a number in the valley in-between the two, such as 375

- If the measured gain is more than 375, it's probably a Zetex transistor.
- If the measured gain is less than 375, it's probably a 3904 transistor.

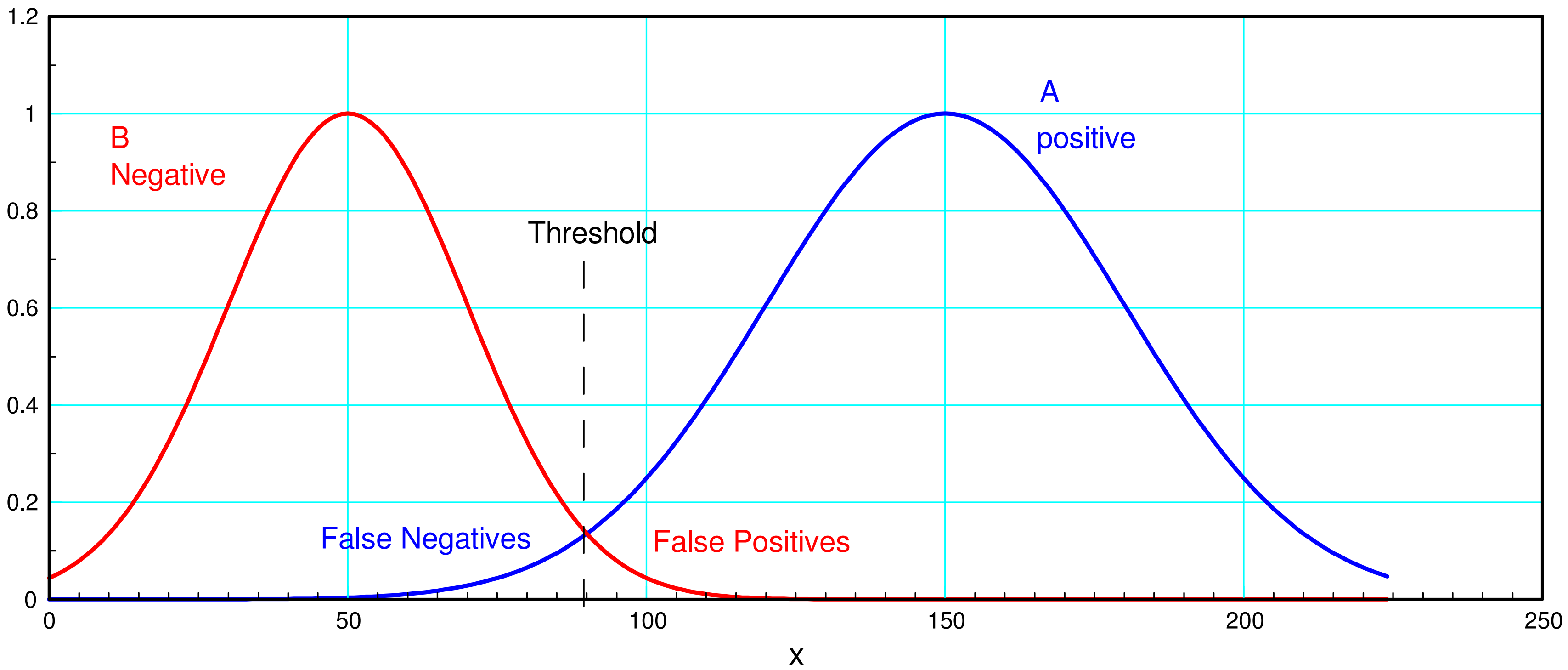
The above is an idea case where the two populations are very distinct. A more common situation is when the two distributions overlap more.

Example 2: Let population A have (positive)

- mean = 150 standard deviation = 35

Let population B have

- mean = 50 standard deviation = 20



Since there is over lap, you will make mistakes when deciding which population a sample belongs to

- False Positive: You think x comes from population A (positive) when it actually came from B
- False Negative: You think x came from population B (negative) when it actually came from A

There are several variations on how to set the threshold

Case 1: p(False Negative) = 1%

For the above example, determine

- The threshold for separating positive (A) and negative (B) results so that the probability of a false negative is 1%.
- With this threshold, determine the probability of a false positive

Solution: Use a standard normal table (or StatTrek) to determine the z-score which corresponds to a tail of 1%

answer: $z = 2.236$

The threshold should then be

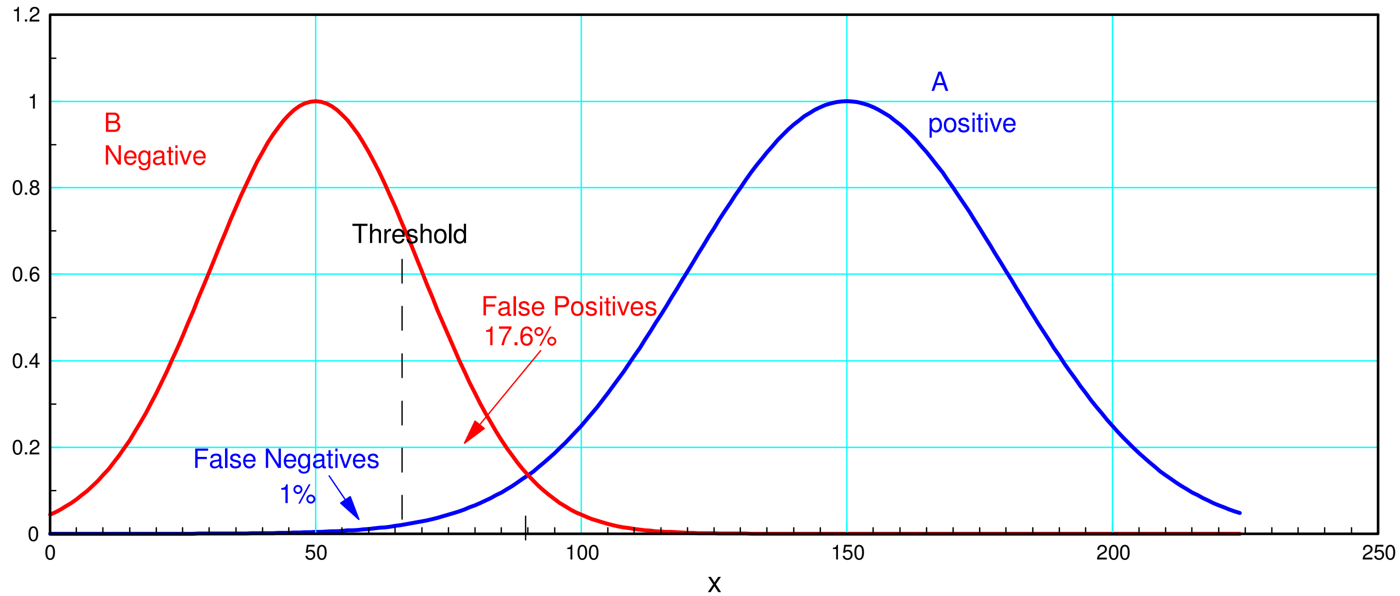
$$T = \mu_A - 2.326\, \sigma_A = 150 - 2.326 \cdot 35 = 68.59$$

The probability of a false positive is found from determining the area of the right-hand tail of population B

$$z = \left(\frac{68.59 - \bar{x}_B}{s_B}\right) = \left(\frac{68.59 - 50}{20}\right) = 0.9295$$

Using a standard normal table to convert this back to a probability:

- p = 0.176



A threshold of 68.59 results in 1% false positives and 17.6% false negatives.

Case 2: p(False Positive) = 1%

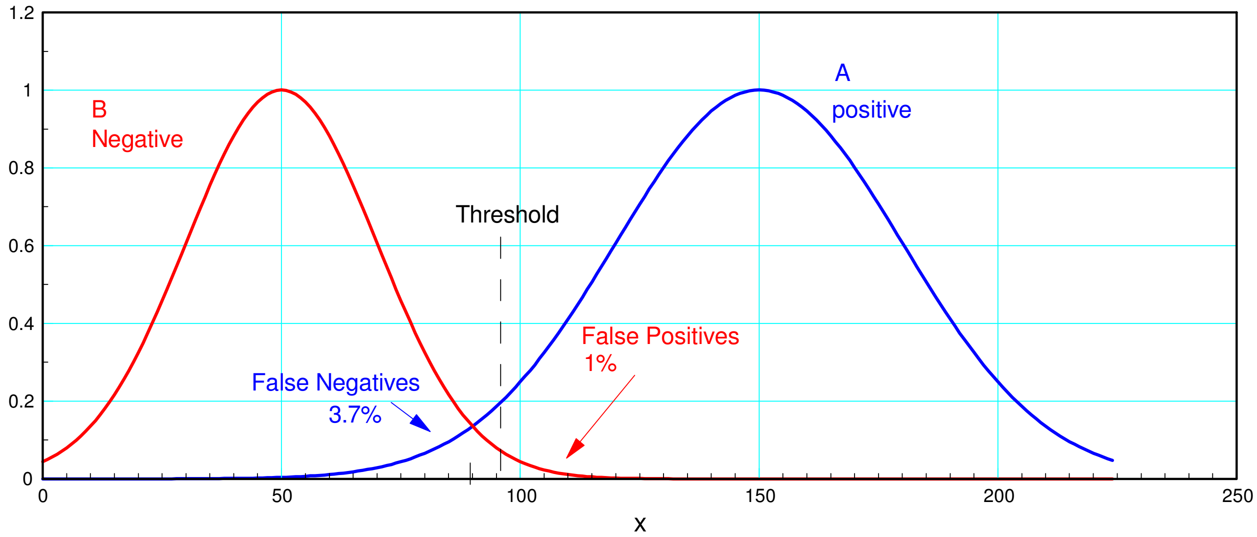
Instead, if you set the probability of a false positive, the threshold should be

$$T = \mu_B + 2.326\,\sigma_B = 50 + 2.326 \cdot 20 = 96.52$$

Now, the probability of a false negative is

$$z = \left(\frac{\bar{x}_A - 96.52}{s_A}\right) = \left(\frac{150 - 96.52}{35}\right) = 1.7827$$

$$p = 0.037$$



Case 3: p(False Positive) = p(False Negative)

For this to be the case, the z-score for a false positive should equal the z-score for a false negative

$$z = \left(\frac{\mu_A - T}{\sigma_A}\right) = \left(\frac{T - \mu_B}{\sigma_B}\right)$$

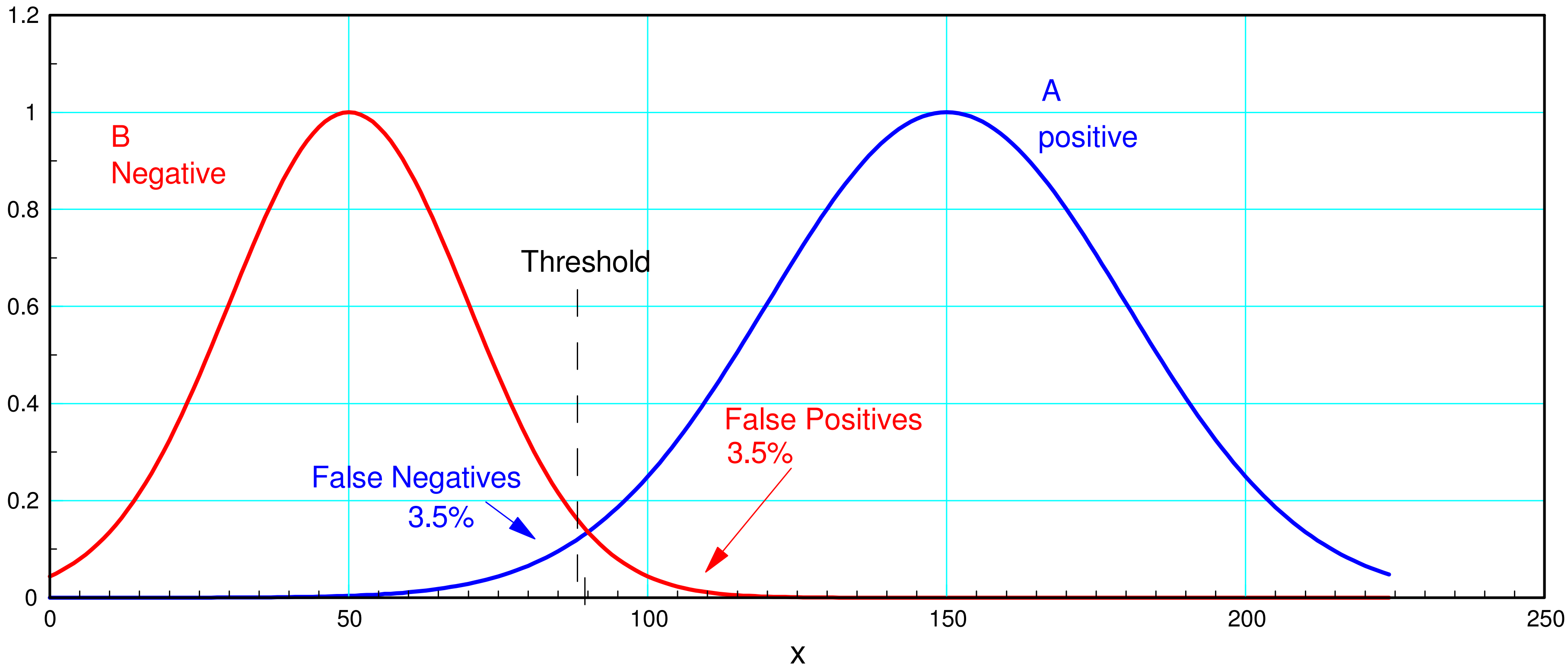
$$\left(\frac{150 - T}{35}\right) = \left(\frac{T - 50}{20}\right)$$

$$T = \left(\frac{20 \cdot 150 + 50 \cdot 35}{20 + 35}\right) = 86.36$$

The probability of a false positive or false negative is then p = 0.035

$$z = \left(\frac{86.36 - 50}{20}\right) = \left(\frac{150 - 86.36}{35}\right) = 1.81812$$

$$p = 0.035$$



A threshold of 86.6 results in the probability of a false positive and false negative being 3.5%

If you want to be more certain, you could run a separate (independent) test. If this test also has a 3.5% error rate, then the two tests will give

- $p = (0.965) (0.965) = 0.9312$ both tests will give the correct result
- $p = (0.035) (0.035) = 0.0012$ both tests will give the incorrect result
- $p = 0.0676$ both tests give different results

With two tests, you can greatly reduce the probability of a false positive or false negative.