Biostatistics

Week #10 4/26/2022

Ch 9 – Confidence Intervals – Part 1

Outline

- 9.0 Introduction

- 9.1 Two-sided Confidence Intervals

- 9.2 One-sided Confidence Intervals (some MATLAB functions for normal distributions)

- 9.3 Student's t Distribution

- 9.4 Applications

9.0 Introduction

- Now we have a *sample* of n limited observations.

- It has a mean value $\overline{x}$ as well as a standard deviation s, computed based on these n observations.

- Can we **estimate** the population statistics, for example, **a population mean** μ, using the information contained in this sample of n observations?

Method 1 - Point Estimation

- Using the sample data to calculate a single number to estimate the parameter of interest. That is, using sample mean $\bar{x}$ to estimate the population mean μ.

- The problem is apparent – two samples might give very different mean. (Uncertainty involved.)

- It does not provide any information about the **inherent variability** of the sample means, nor about **the sample size**.

Method 2 – Interval Estimation

- Provide **a range of reasonable values** that are **intended** to contain the parameter of interest – the population mean μ in this case, **with a certain degree of confidence**.

- This range is called a **confidence interval, or CI**.

- "I am confident that the population mean μ should be included in this interval."

# 9.1 Two-sided Confidence Intervals for $\bar{x}$

- From previous lecture we learned that **the sampling distribution of the mean** is a normal distribution.

- Given a random variable $\bar{x}$ representing many sample means, and the population has mean μ and standard deviation σ, we know that the following conversion leads to a standard normal distribution for Z:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Note that the random variable used here is the sampling means. Each sample consists of n individuals, with a population mean value μ and population standard deviation σ .
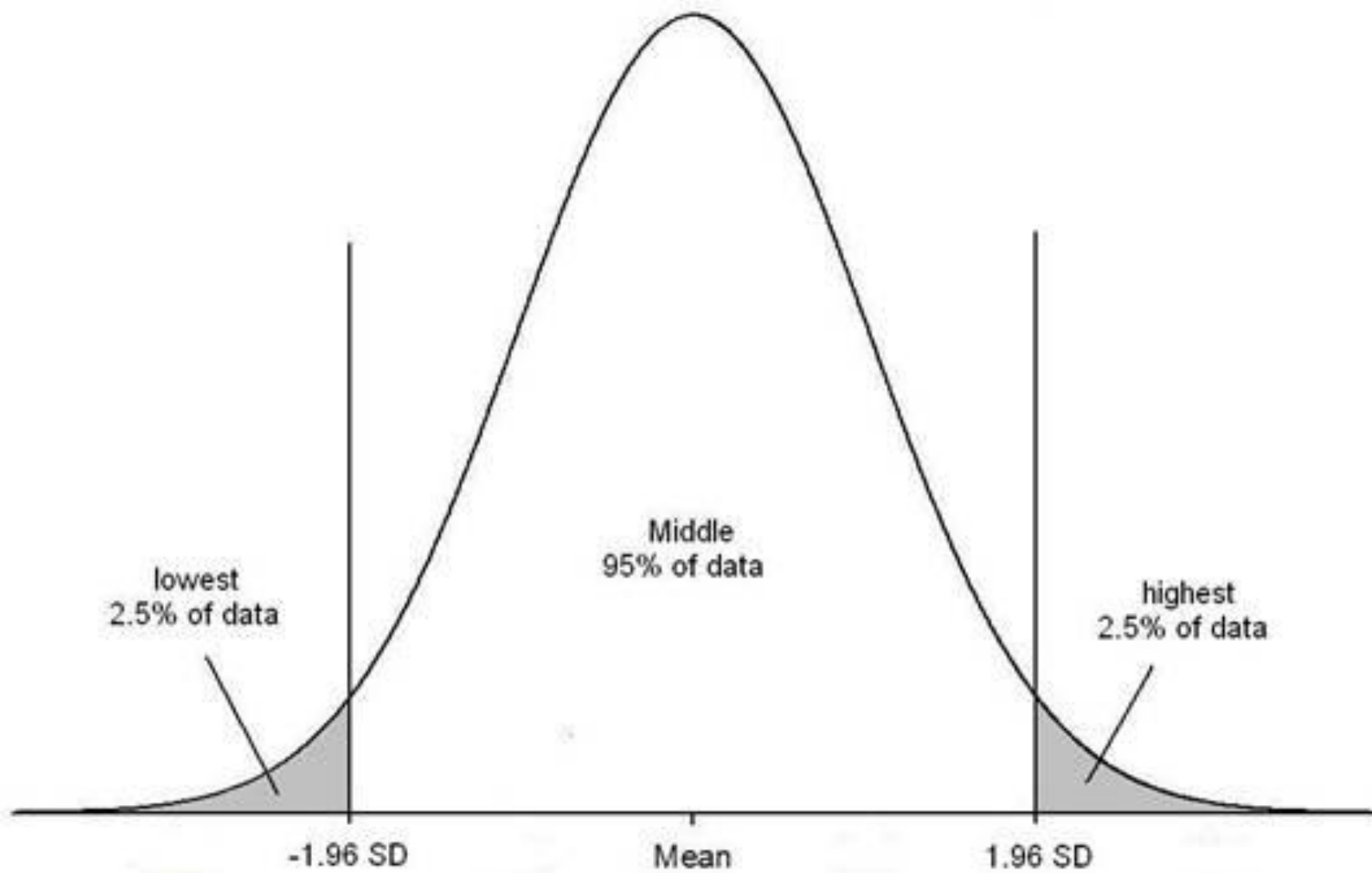
Tables of the Normal Distribution
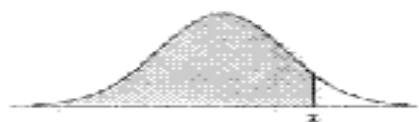
Probability Content from −oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834						0.9857
2.2	0.9861	0.9864	0.9868	0.9871					7	0.9890
2.3	0.9893	0.9896	0.9898	0.9901					3	0.9916
2.4	0.9918	0.9920	0.9922	0.9925					4	0.9936
2.5	0.9938	0.9940	0.9941	0.9943					1	0.9952
2.6	0.9953	0.9955	0.9956	0.9957					3	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

This gives a probability of 0.9750 for z = -inf to 1.96, based on a standard normal distribution.

- We know that 95% of the sample means, after converted to Z, will lie between Z=-1.96 to Z=1.96. That is:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \leq 1.96) = 0.95$$

$$P(-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(-1.96 \frac{\sigma}{\sqrt{n}} - \bar{x} \leq -\mu \leq 1.96 \frac{\sigma}{\sqrt{n}} - \bar{x}) = 0.95$$

$$P(\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(\boxed{\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}} \leq \mu \leq \boxed{\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}}) = 0.95$$

The quantities in red boxes are the boundaries for the **95% confidence interval**.

95% Confidence Interval

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

- 95% boundaries of the population mean; **we are 95% confident that the interval will contain μ.**

- It is **NOT** saying that μ is a random variable that takes a value within the interval 95% of the time.

- It is **NOT** saying that 95% of the population mean values lie between these boundaries.

95% Confidence Interval

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

- It means that if we were to select **100 random samples** from the population and use these samples to calculate **100 different confidence intervals for** μ, approximately **95 of the intervals would cover the true population mean** μ and 5 would not.

95% Confidence Interval

$\alpha = 0.05$

$Z_{\alpha/2} = 1.96$

95%

z = −1.96

z = +1.96
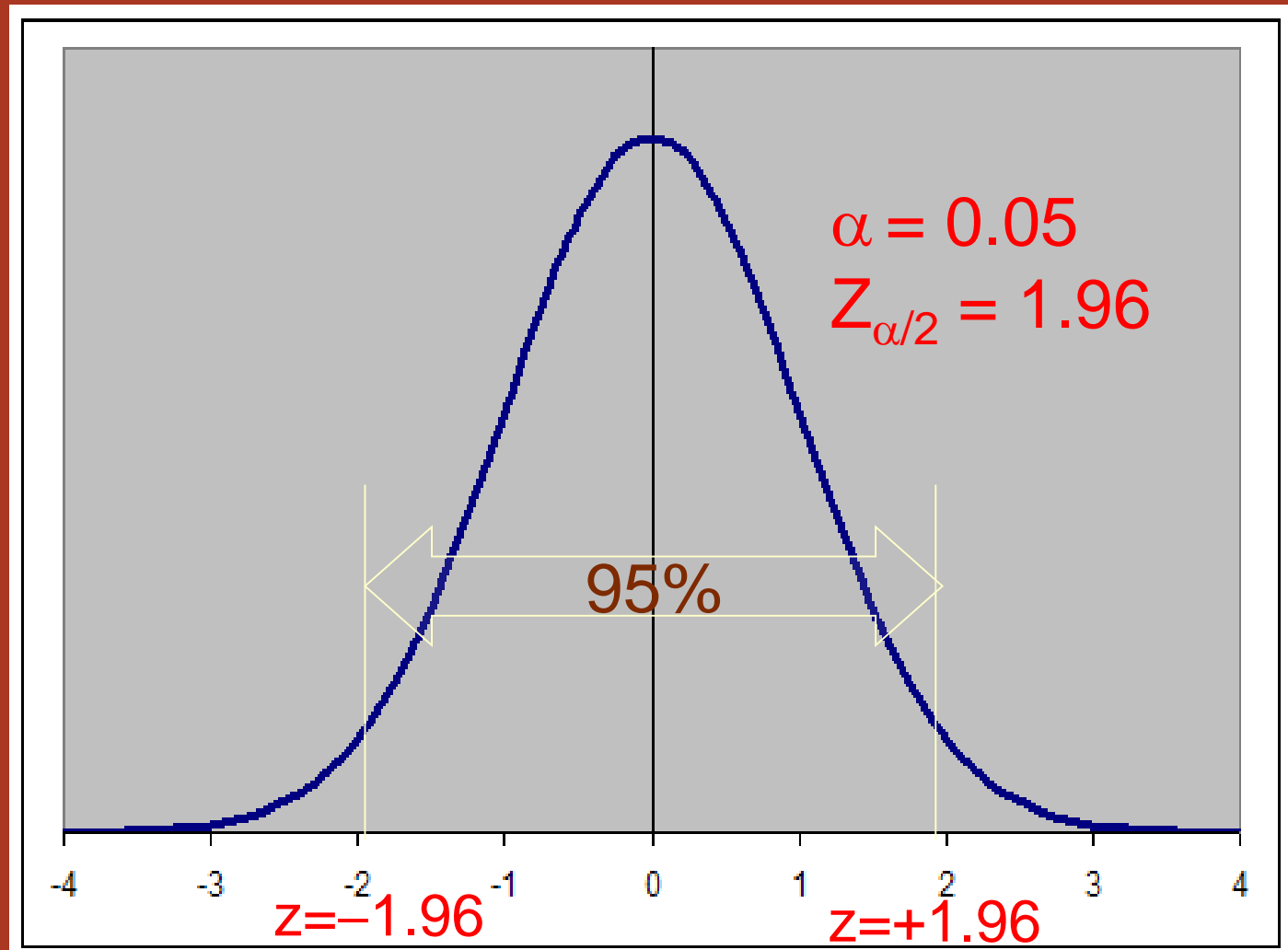
99% Confidence Interval

$$(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}})$$

- This is to say that Z from -2.58 to 2.58 covers 99% of the area under curve for a standard normal distribution.

99% Confidence Interval

$\alpha = 0.01$

$Z_{\alpha/2} = 2.58$

99%

z=−2.58

z=+2.58
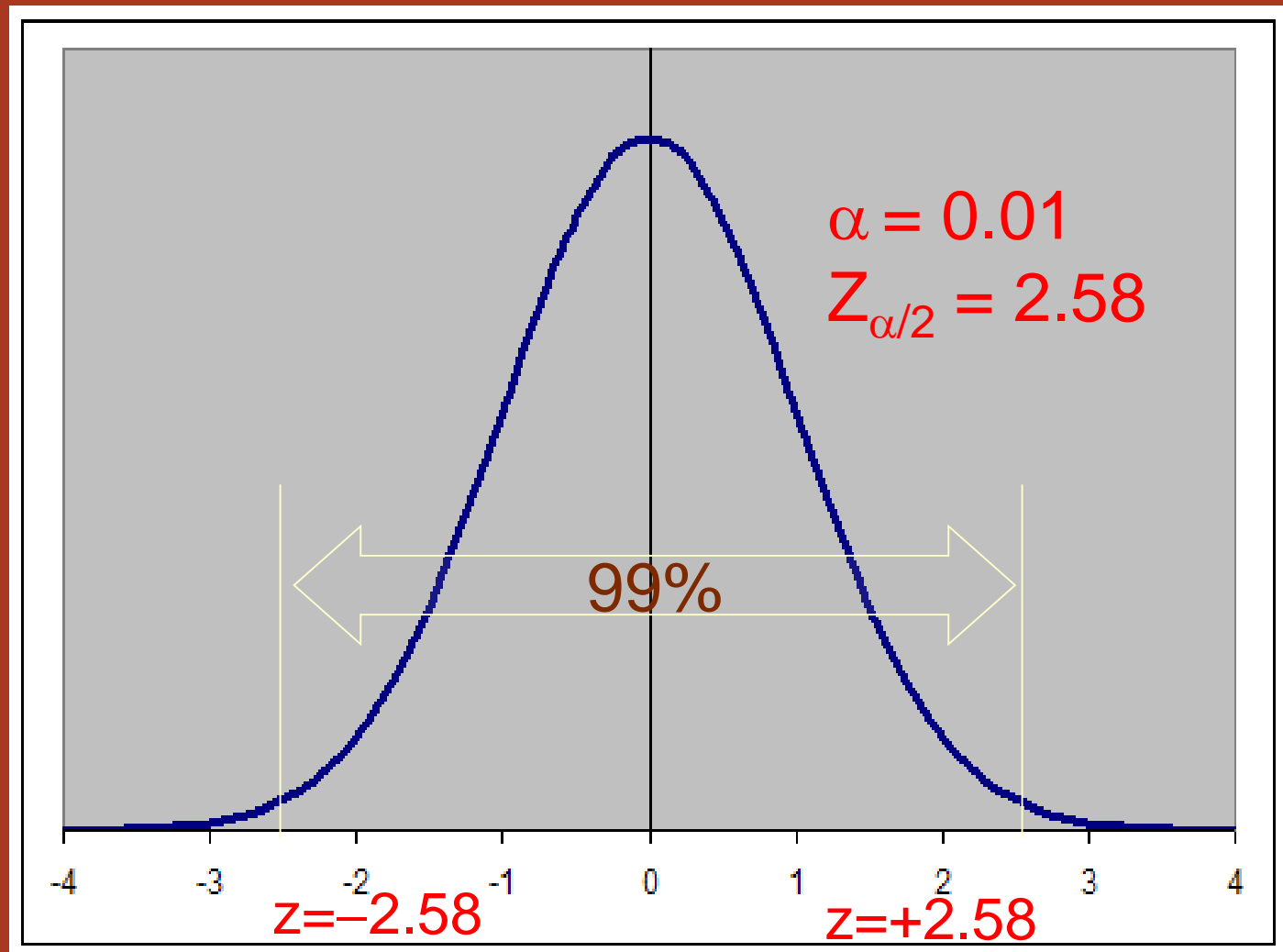
Example 1

- Consider the distribution of serum cholesterol levels for all males in US who are *hypertensive* and who *smoke*.

- The distribution is approximately normal with **an unknown mean** μ and standard deviation $\sigma=46$ mg/100 ml.

- We are interested in estimating a mean serum cholesterol level of **this** population.

Example 1 (cont'd)

- Before we go out and select a random sample, the probability that this interval

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

covers the true population mean μ is **0.95**.

- Taking n=12 and assuming that the mean value computed from these 12 individuals is **217**.

Example 1 (cont'd)

We may calculate this interval:

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

$$(217 - 1.96\frac{46}{\sqrt{12}}, 217 + 1.96\frac{46}{\sqrt{12}})$$

$$(217 - 26.027, 217 + 26.027)$$

$$(191, 243)$$

Thus the 95% confidence interval is (191, 243).

Example 1 (cont'd)

- What does this mean?
- While 217 (the computed mean from these 12 individuals) is our best guess for the mean value from the population, the interval of 191 to 243 provides a range of **reasonable** values for the population mean μ.
- We are 95% confident that the limits 191 and 243 cover the true mean μ.

Example 1 (cont'd)

- We DO NOT say that there is a 95% chance that the μ lies between these values; μ is fixed and either it is between 191 and 243 or it is not.

Example 1 (cont'd)

Numerical simulation of 100 random samples of n=12 from this population. Each sample computes for its own CI, and all these CIs are of the same length. Indeed there are only 5 of them did not include *the actual population mean value 211* (the horizontal line).

Example 1 – cont'd

- Note that the 95% CI covers from 191 to 243, or a range width of 52.

- Instead of 95% CI, we may get the 99% CI by changing Z=1.96 to 2.58.

- This gives (183, 251), or a width of a **wider** range 68.
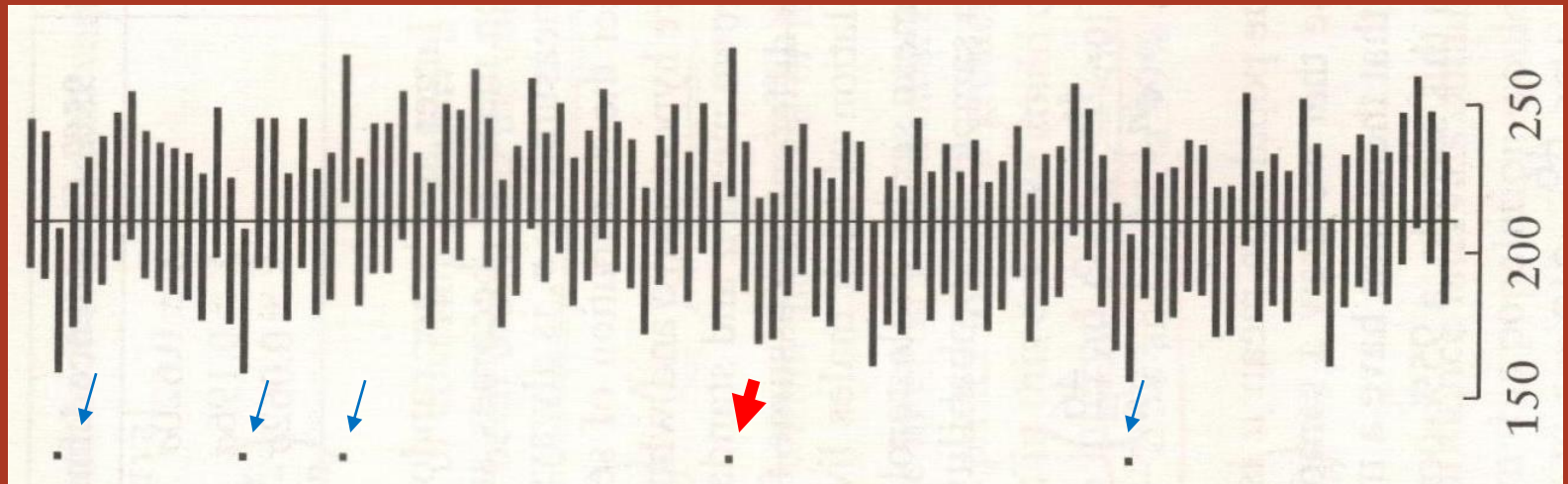
It is reasonable that, when CI gets wider, at least there will be some (blue arrows) of the 5 CIs that previously excluded the population mean 211 now covering that mean value. Looks like one (red arrow) won't contain 211 even with a wider CI range of 68.

Example 2

- In Example 1, the length of the 99% CI interval gets larger than a 95% CI, from 52 to 68.

- The length of CI gets larger when the level of confidence Z gets bigger.

- **In fact, the length of CI narrows when σ gets smaller or n gets bigger.**

$$\left(\bar{x} - Z \frac{\sigma}{\sqrt{n}}, \bar{x} + Z \frac{\sigma}{\sqrt{n}} \right)$$

Z=1.96 for 95% CI
Z=2.58 for 99% CI

Example 2 – cont'd

- **How large a sample size would be to reduce the length of 99% CI to only 20?**

- Recall that the interval is centered at 217. So the lower bound would be 217-10=207 and upper bound be 217+10=227.

$$(217 - 2.58\frac{46}{\sqrt{n}}, \ 217 + 2.58\frac{46}{\sqrt{n}})$$

or $2.58\frac{46}{\sqrt{n}} = 10$ This gives n =140.8

Notes on Confidence Intervals

- **Interpretation**
  - Possible values for the population mean μ with high confidence
- **Are all CIs 95%?**
  - No
  - It is the most commonly used
  - A 99% CI is wider
  - A 90% CI is narrower

Cont'd

- **Random sampling error**
  - Confidence interval only accounts for **<u>random sampling</u>** error—not other systematic sources of error or bias
- **Examples of <u>Systematic Bias</u>**
  - Blood Pressure (BP) measurement is always +5 too high (broken instrument)
  - Only those with high BP agree to participate (non-response bias)

Is CI Wider good or bad?

- A wider interval means that there exists **_bigger variation_** among sample means.
- This could be due to bigger σ, smaller sample size n, or bigger Z to use.

$$(\bar{x} - Z\frac{\sigma}{\sqrt{n}}, \bar{x} + Z\frac{\sigma}{\sqrt{n}})$$

Cont'd

- This poses uncertainties that whether a sample is good enough to represent the population mean.

- To get a reliable sample, however, **we desire a high level of confidence.** (For example, we want 95% rather than 90%.)

- This results in a wider CI (uncertainty).

- To compensate the widening of the interval, we **need to increase n** if the population variation cannot be overlooked.

Example 3

- Blood pressure
  $n = 100$, $\bar{x} = 125$ mm Hg, $\sigma = 14$

- We know that the CI is defined as:

$$(\bar{x} - Z\frac{\sigma}{\sqrt{n}}, \bar{x} + Z\frac{\sigma}{\sqrt{n}})$$

- Therefore we firstly compute

$$\frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{100}} = 1.4$$

Example 3 – cont'd

95% CI for μ (mean blood pressure in the population) uses Z=1.96. Therefore the CI becomes:

$$125 \pm 1.96 \times 1.4$$

or

125 ± 2.744

$$(\bar{x} - Z \frac{\sigma}{\sqrt{n}}, \bar{x} + Z \frac{\sigma}{\sqrt{n}})$$

This is the interval to find my population mean.

Ways to Write a Confidence Interval

- 122.2 to 127.8 (length=5.6)

- (122.2, 127.8)

- 122.2–127.8 ⟵ (122.2 **to** 127.8, not 122.2 minus 127.8)

- The **95% error bound** on $\overline{x}$ is 2.8

The mean value is 125. The variation is ±2.8mm, or 2.8/125= ±2.24%, at a confidence level of 95%.

Underlying Assumptions for 95% CI

- In order to be able to use the formula

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

  the data must meet a few conditions that satisfy the underlying assumptions necessary to use this result

- *Assumptions:*

  – Random sample of population—important!

  – Observations is sample **independent**

  – Sample size n is **at least 30~60** (we will explain later) [Central limit theorem requires large n!]

t-correction

- If sample size is smaller than 30
  - The sampling distribution of the means is **not** quite normally distributed

  - It instead approximates a "*t-distribution*" (which we will talk about later in subsequent lectures)

  - There needs a small correction — called the ***t-correction***

  - That is, the number 1.96 in the formula below needs to get slightly **bigger** (to achieve the same 95% confidence level)

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

Two-sided vs one-sided CI?

- Also known as two-*tailed* or one-*tailed* (because of "tails" in a bell-shaped distribution)

- It depends on whether only one direction is considered extreme (and unlikely) or both directions are considered extreme.

9.2 One-sided Confidence Interval

- In some situations, we are concerned with either an upper limit or a lower limit for μ, but not both. (Only one direction is considered 'extreme' or 'not likely'.)

- In this case, we consider only **one-sided** instead of two-sided CI.

- Recall that in a 2-sided case, we have **Z-value between –1.96 and +1.96** to cover 95% for a standard normal distribution.

Cont'd

- For one-sided, we consider either {$-\infty$, 1.96} or {-1.96, ∞} as normal. This would cover, apparently, **more than 95%. (In fact, 97.5%)**
- To cover only 95%, this Z-value (absolute value) should be **smaller**.
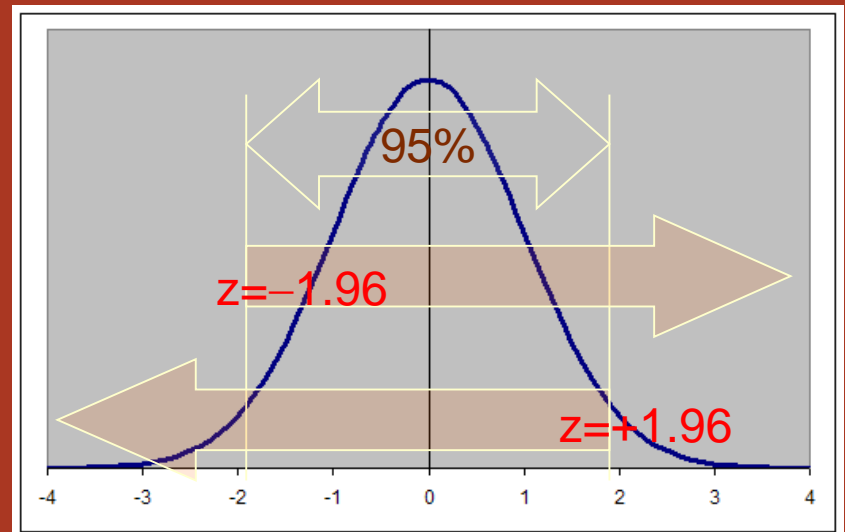
A Case of Left-tailed

- This problem translates into:
  - What is the value for z for $P(Z \geq z) = 0.95$, for a standard normal distribution.
  - ☑ **Answer = –1.645**
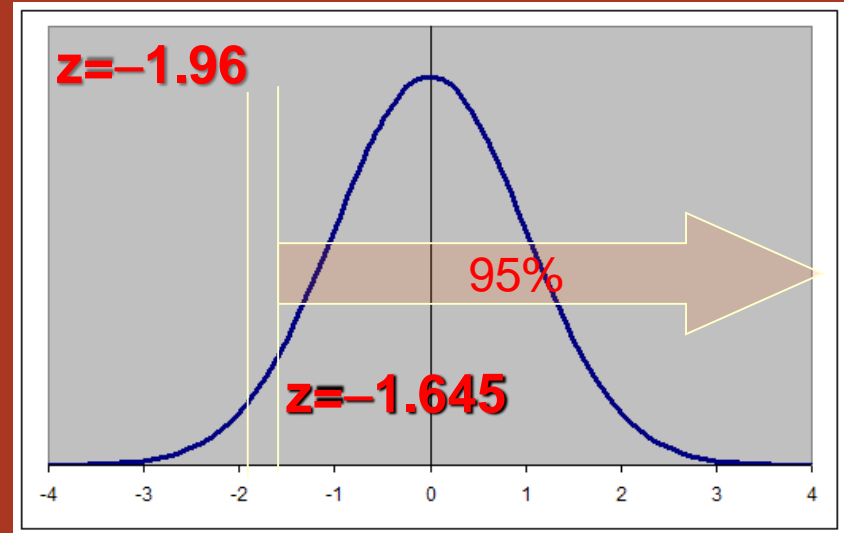
```
>> F='1/(sqrt(2*pi))*exp(-0.5*z^2)';
>> z=?
>> int(F,z,inf)
ans =.95
>>
```

Texts shown on the left are not actual MATLAB commands. What we need to know is "What value of z would give the integration a result of 0.95".

Example #4

- Consider a distribution for hemoglobin (血紅素) levels for US children < 6 years old who have been exposed to high level of leads (鉛) (**thus have lower hemoglobin levels**). 'Low' is bad!!!

- This distribution has an **unknown mean value** μ and $\sigma = 0.85$ g/100 ml.

- We are interested in knowing **the upper bound for** μ. (So that if your hemoglobin level is **lower** than this value, you might be subject to lead poisoning.)

Cont'd

- Recall that the Z-transform for **converting the sampling distribution** into a standard normal distribution is

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

Note that we are converting the "sampling distribution" (that's why we have the sample size *n* here), not the "probability distribution" of the random variable *X*.

Standard normal distribution of hemoglobin (血紅素) levels for 74 children who have been exposed to high level of leads. This sample mean is 10.6 (X=10.6) and knowing that σ = 0.85.
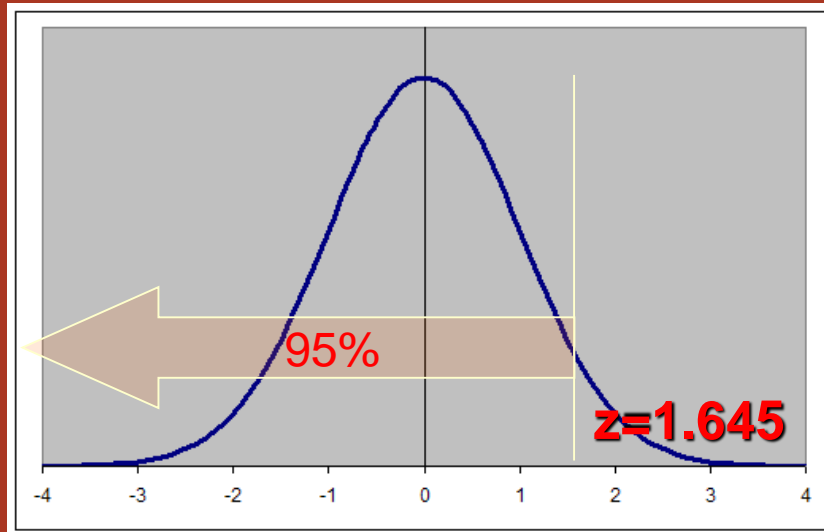
We knew from earlier tries that Z=1.645 can be used for giving this 0.95 probability.

- The actual population mean μ could be covered at most to

$$10.6 + 1.645 \times \frac{0.85}{\sqrt{74}}$$

$$= 10.6 + 0.163 = 10.763$$

Cont'd

- It shows, although this sampling result (from 74 children) gives a mean value of 10.6, the actual mean could be as large as 10.763 (**we are 95% confident about making this statement**).

- If we were to select 100 random samples of size n=74 and use each one to constructed a one-sided 95% CI, approximately 95 of these CI would contain the true mean μ (although we don't really know what that mean value might be).

Remark

- We may compute the mean value from 74 children who have been exposed to high level of leads and get a value of 10.6. [This is called "Point-Estimation" earlier.]

- With this, we may say one kid's hemoglobin level of 10.5 is poisoned (lower than expected), and one with 10.7 is not.

- With one-sided estimation of CI from this n=74 sample, on the other hand, **we will consider** *both kids poisoned*.

MATLAB normpdf

- NORMPDF Normal probability density function (pdf).

- Y = NORMPDF(X,MU,SIGMA) returns the pdf of the normal distribution with **mean MU** and **standard deviation SIGMA**, evaluated at the values in X.

- The size of Y is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

- **Default values for MU and SIGMA are 0 and 1 respectively**. (This is a standard normal distribution.)

Standard normal distribution (taking default MU=0, SIGMA=1)


```
>> x=[-4:0.1:4];
>> y=normpdf(x);
>> plot(x,y)
```

A normal distribution taking MU=129, SIGMA=19.8)


```
>> x=[69:189];
>> y=normpdf(x, 129, 19.8);
>> plot(x,y)
```
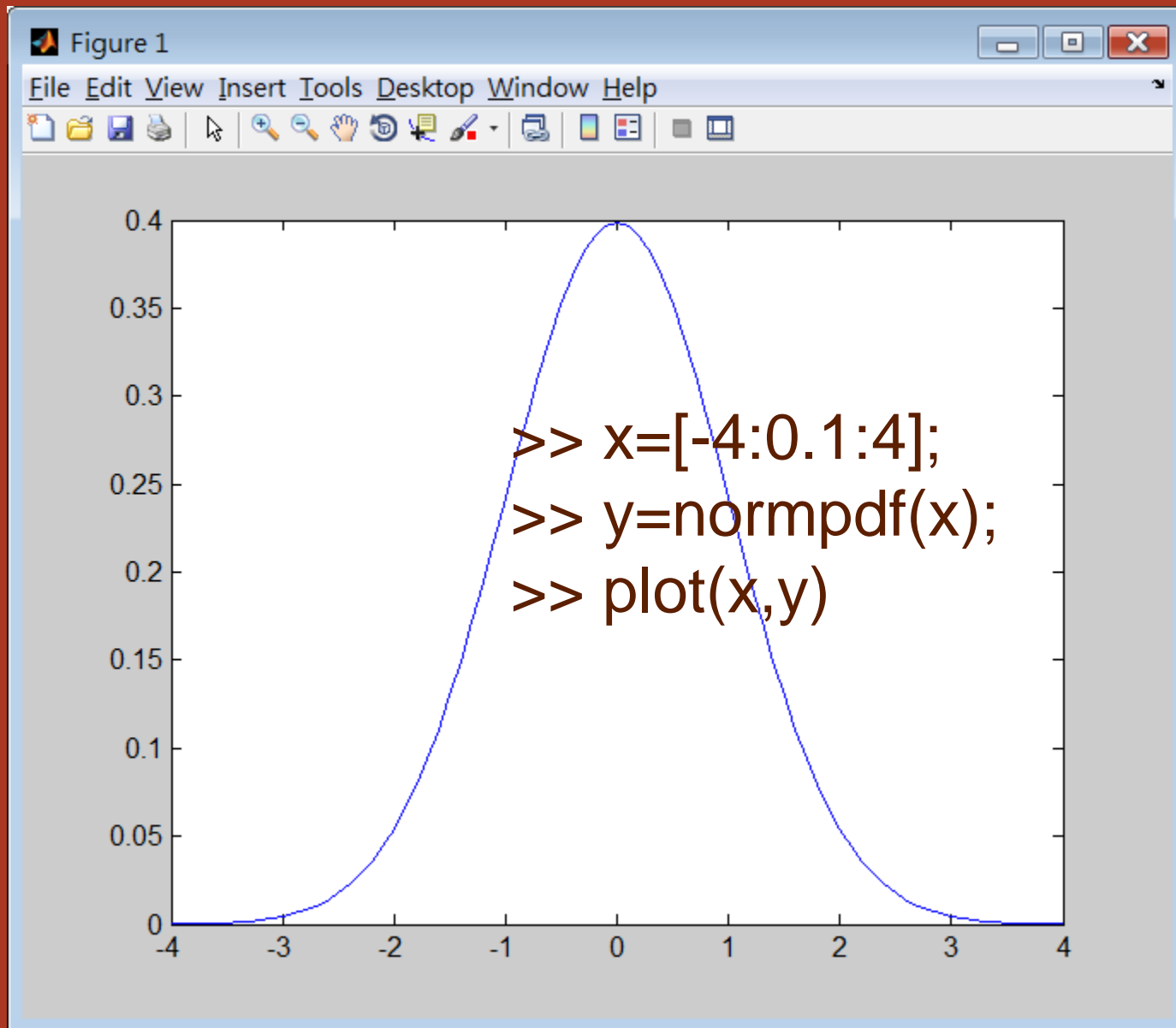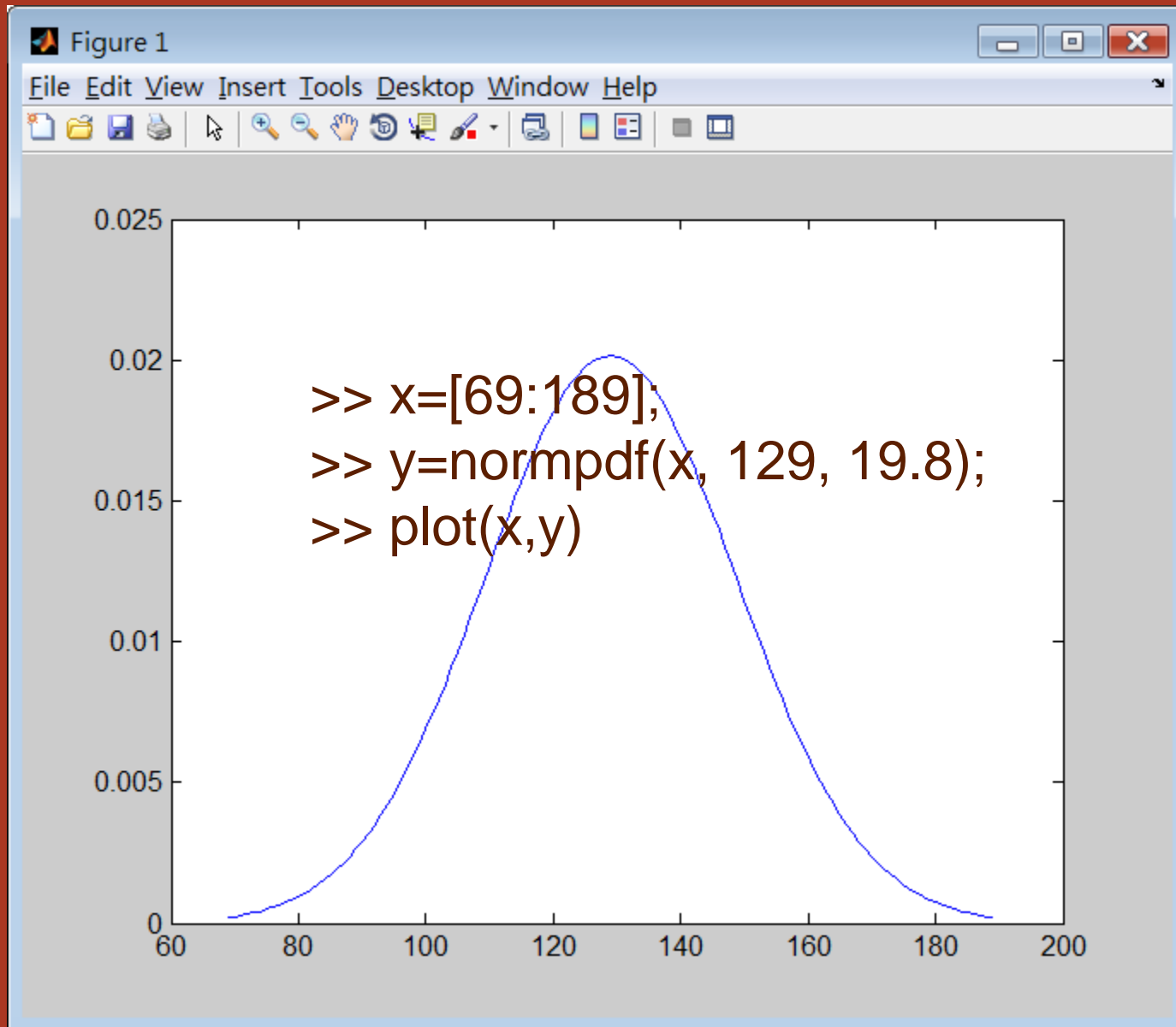
MATLAB normcdf

- NORMCDF Normal **cumulative distribution function** (cdf).

- P = NORMCDF(X,MU,SIGMA) returns the cdf of the normal distribution with mean MU and standard deviation SIGMA, evaluated at the values in X.

- The size of P is the common size of X, MU and SIGMA. A scalar input functions as a constant matrix of the same size as the other inputs.

- Default values for MU and SIGMA are 0 and 1, respectively.

# Standard normal distribution (taking default MU=0, SIGMA=1)
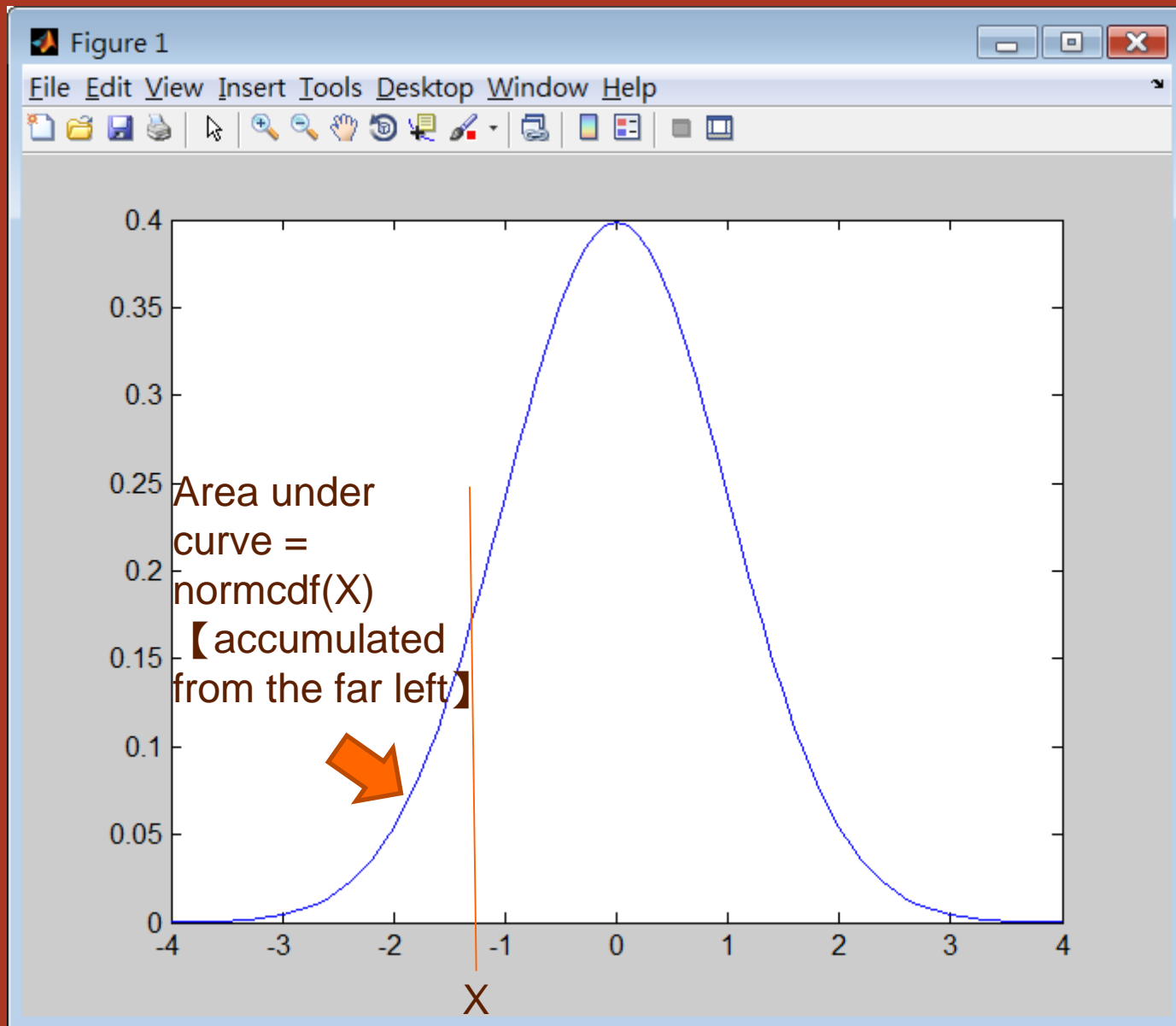
Standard normal distribution (taking default MU=0, SIGMA=1)
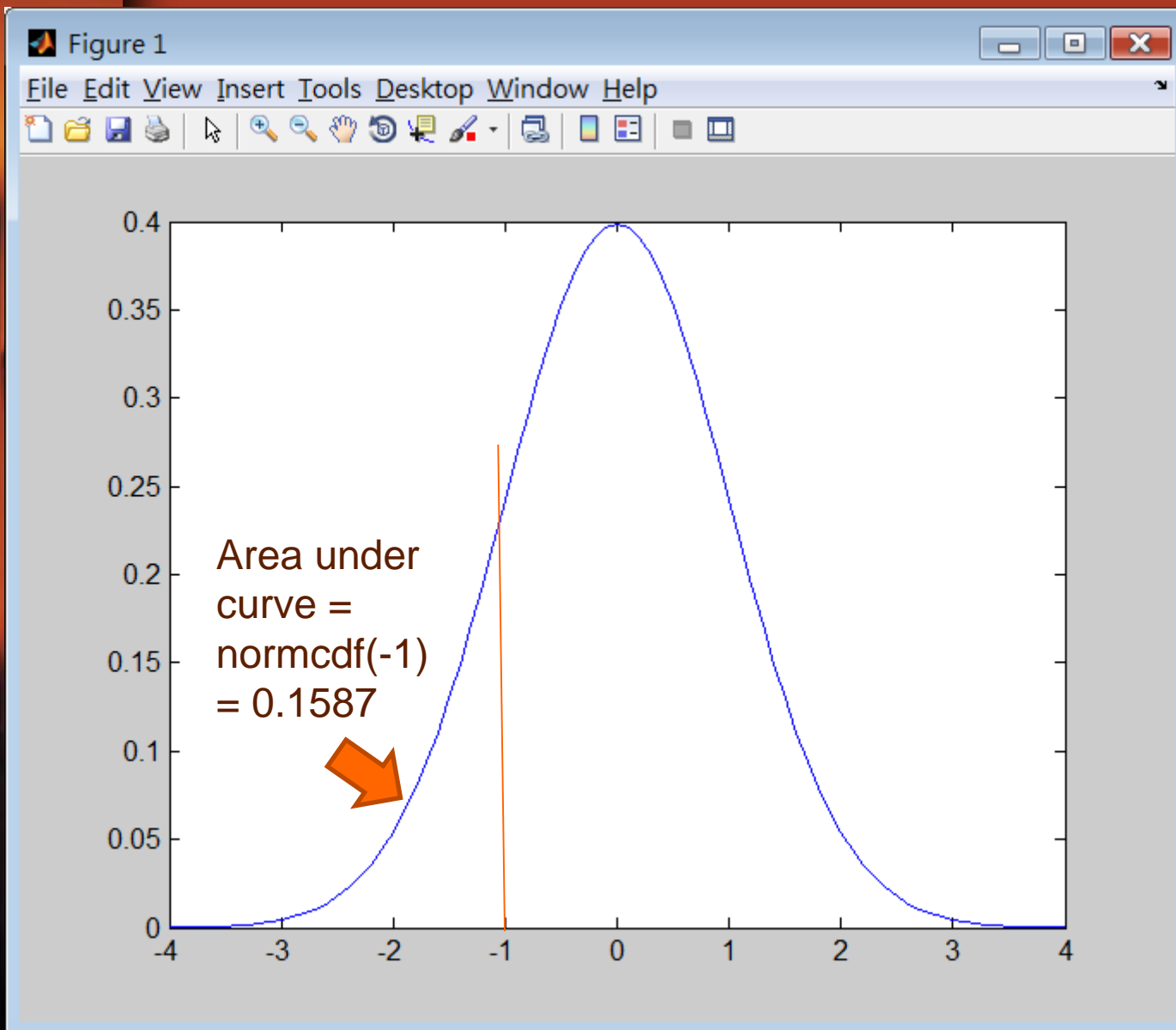

```
>> normcdf(-1)
ans =
    0.1587
>> normcdf(1)
ans =
    0.8413
```

Area under curve = normcdf(-1) = 0.1587

Standard normal distribution (taking default MU=0, SIGMA=1)

Area under curve = normcdf(-2) = 0.0228
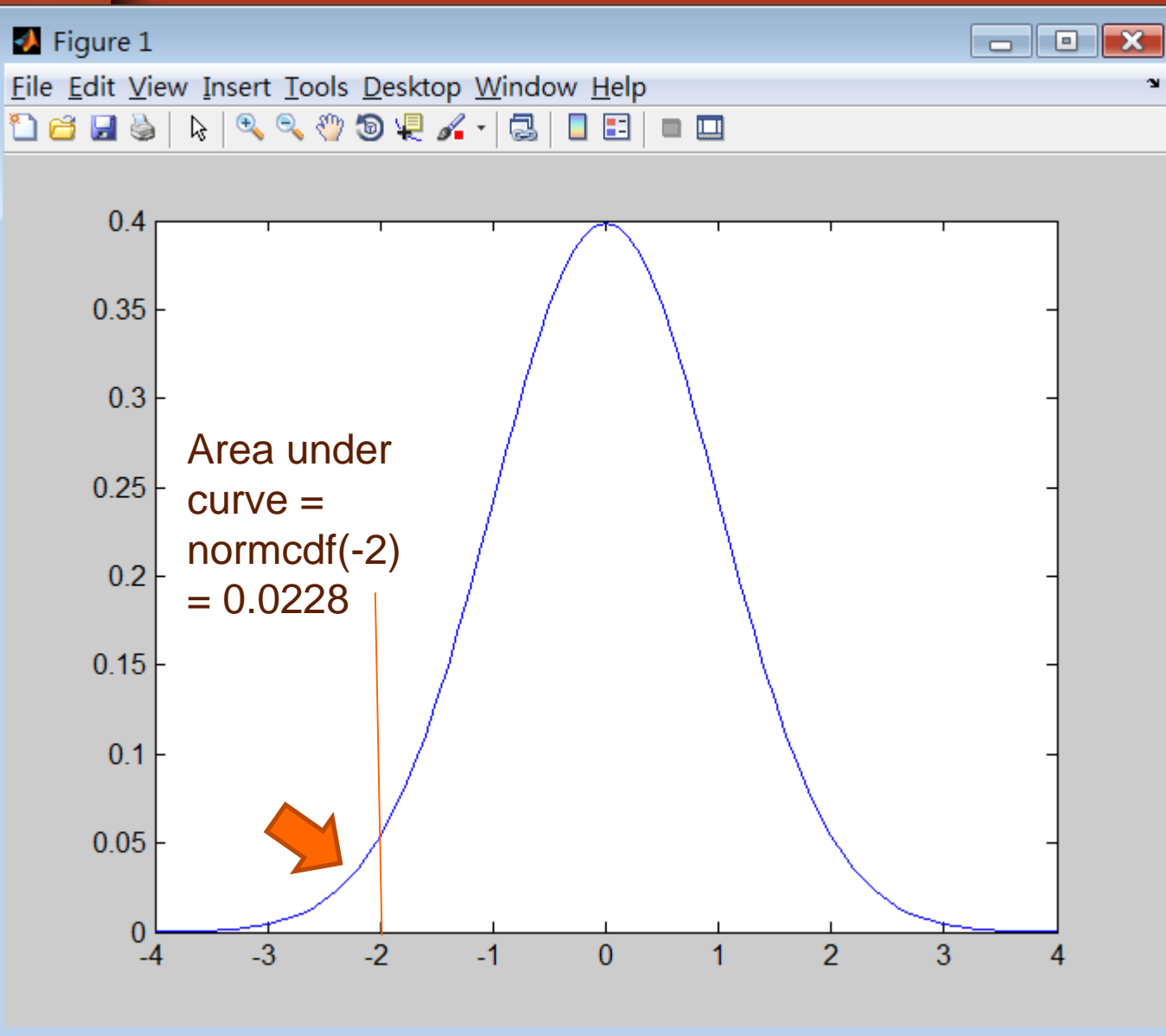
```
>> normcdf(-2)
ans =
    0.0228

>> normcdf(2)
ans =
    0.9772
```

Standard normal distribution (taking default MU=0, SIGMA=1)

What would be the answer for normcdf(0)?

Area under curve = normcdf(-3) = 0.0013
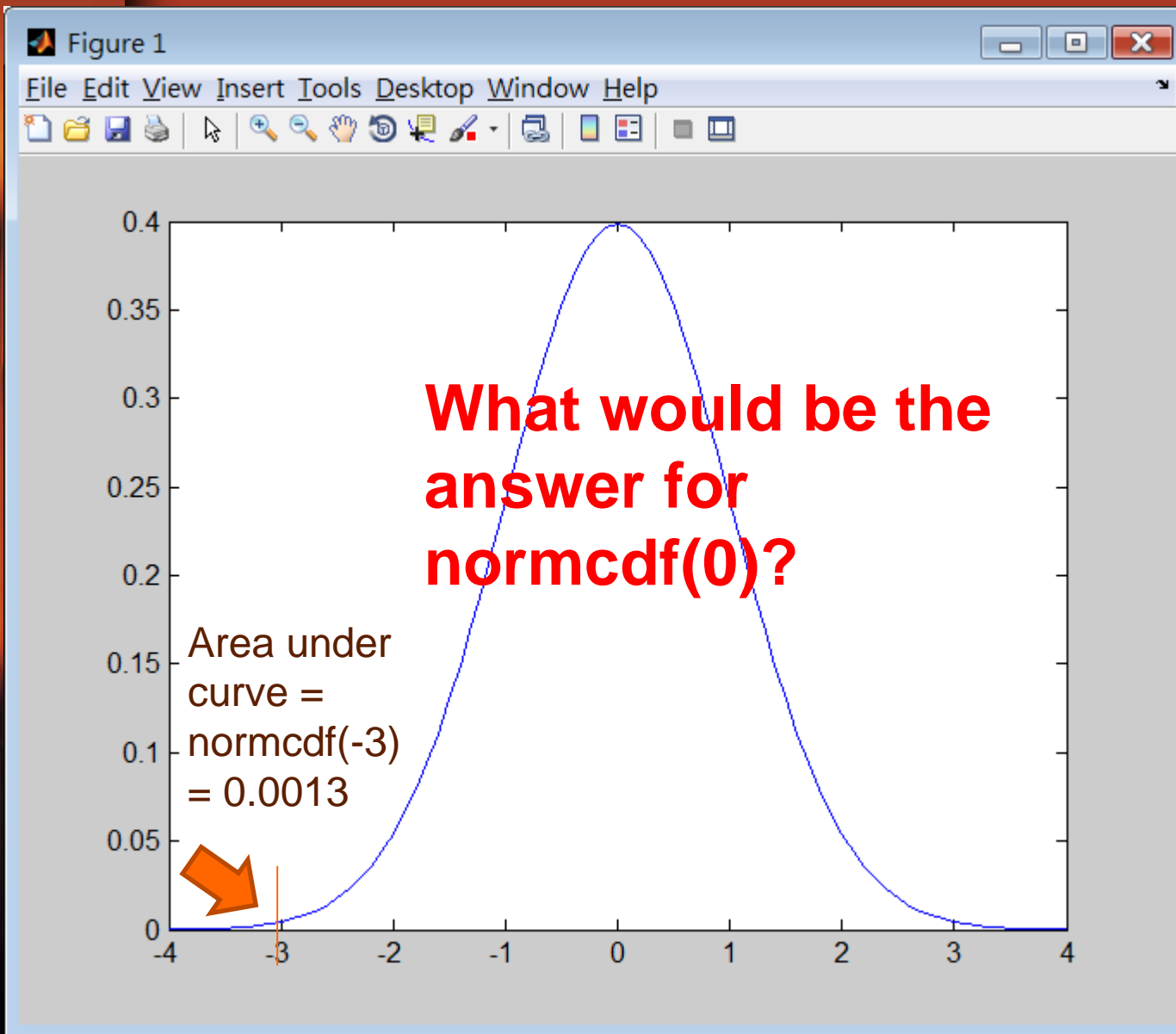
>> normcdf(-3)
ans =
**   0.0013**

>> normcdf(3)
ans =
    0.9987
>>

Standard normal distribution (taking default MU=0, SIGMA=1)
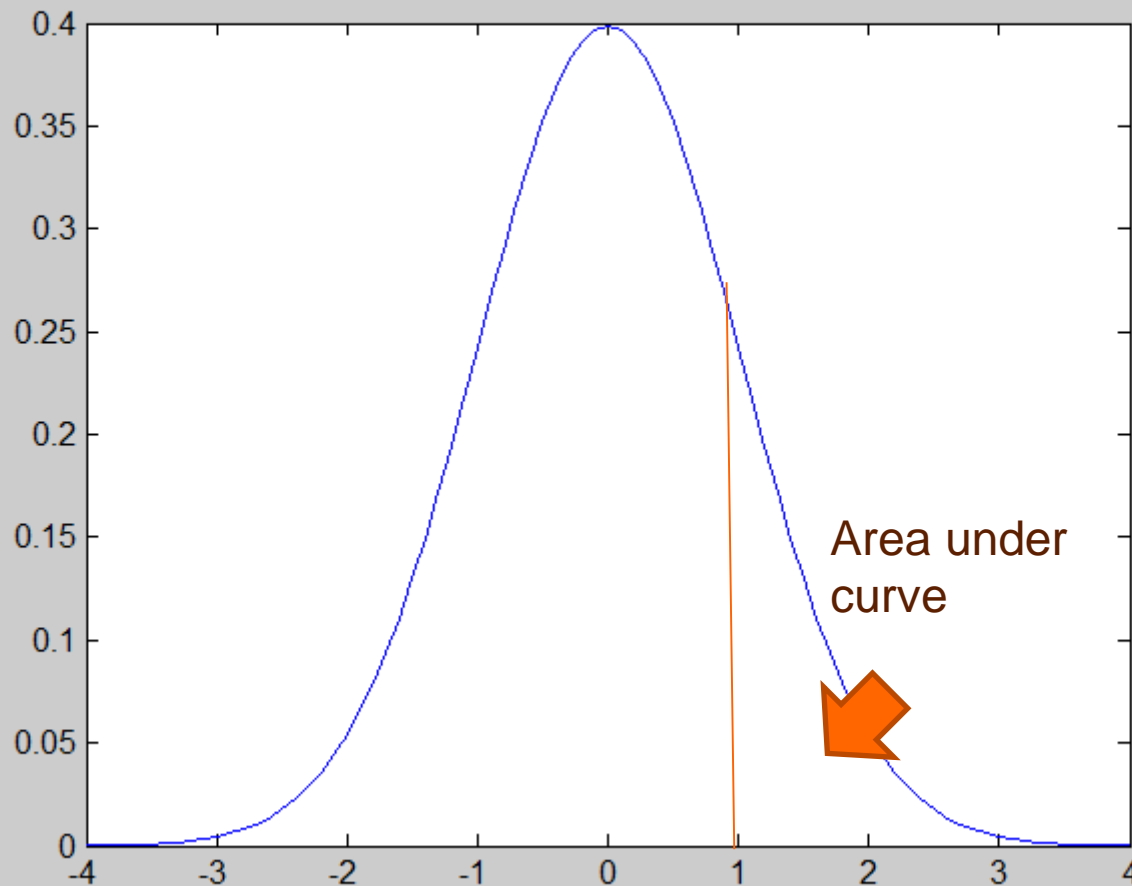
>> 1-normcdf(1)
ans =
    0.1587

A regular normal distribution (taking MU=129, SIGMA=19.8)

A regular normal distribution (taking MU=129, SIGMA=19.8)
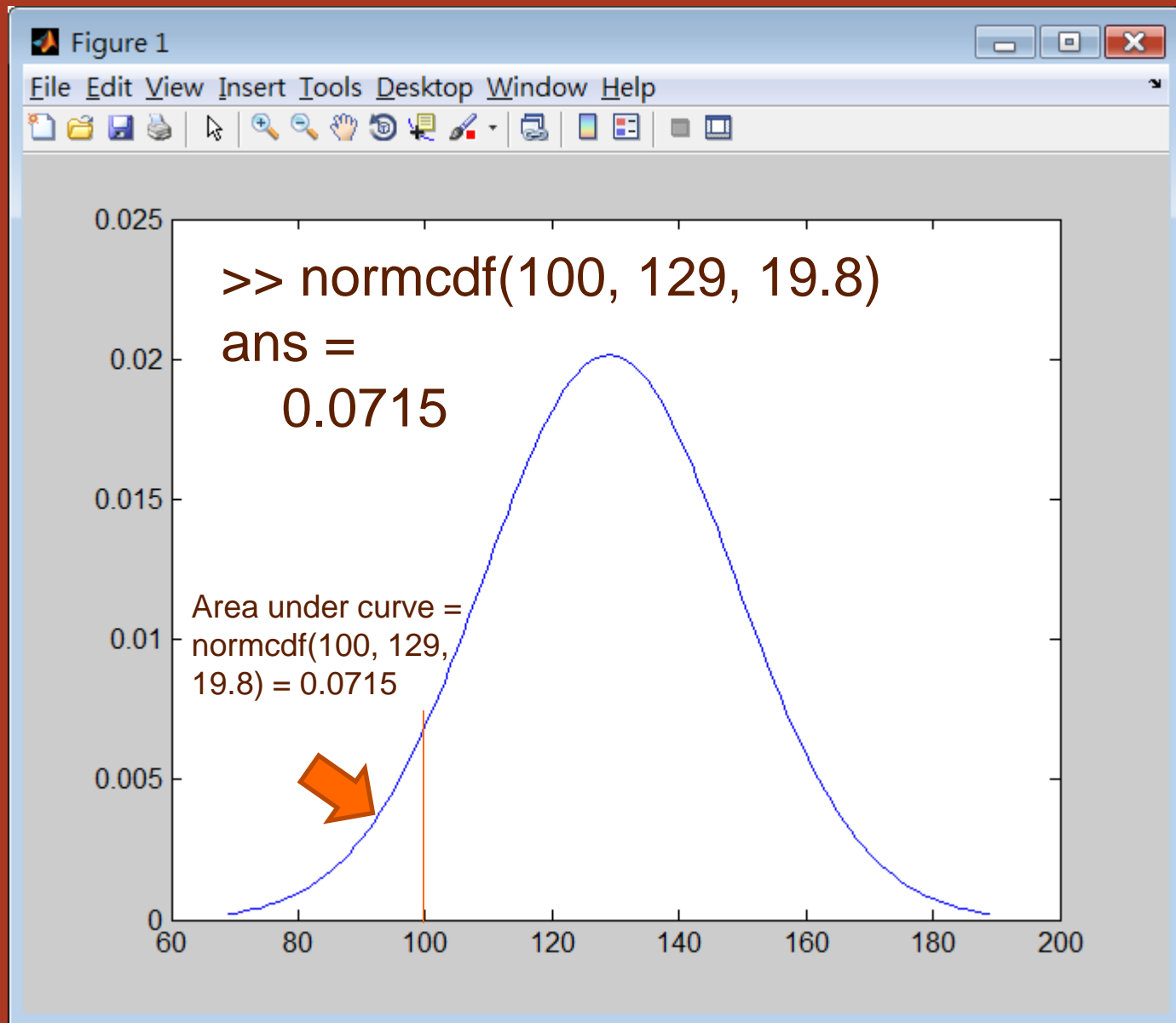
What would be the answer for normcdf(129, 129, 19.8)?
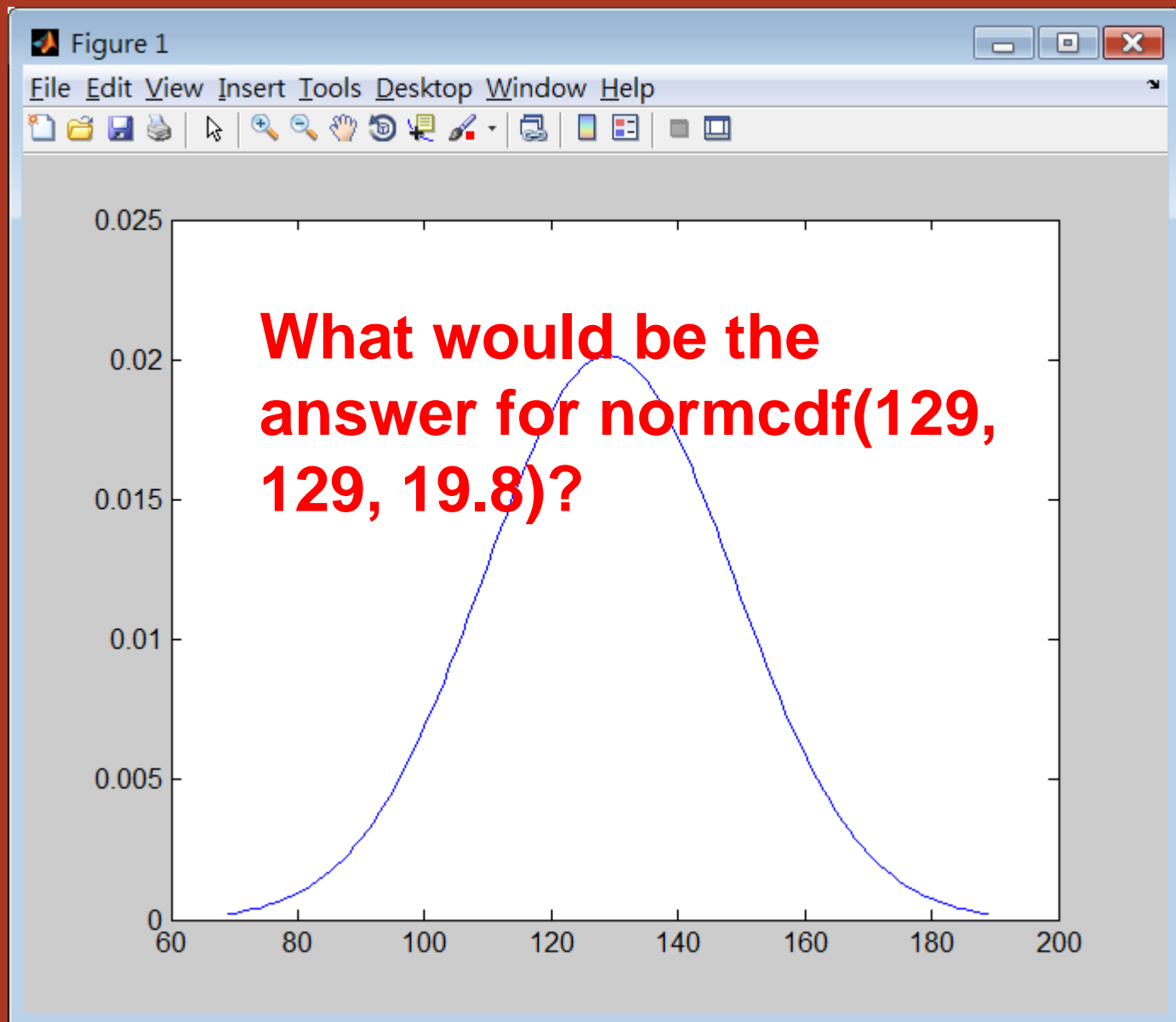
MATLAB norminv

- NORMINV **Inverse of the normal cumulative distribution function** (cdf).

- X = NORMINV(P,MU,SIGMA) returns the inverse cdf for the normal distribution with mean MU and standard deviation SIGMA, evaluated at the values in P.
- The size of X is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.
- Default values for MU and SIGMA are 0 and 1, respectively.

Standard normal distribution (taking default MU=0, SIGMA=1)
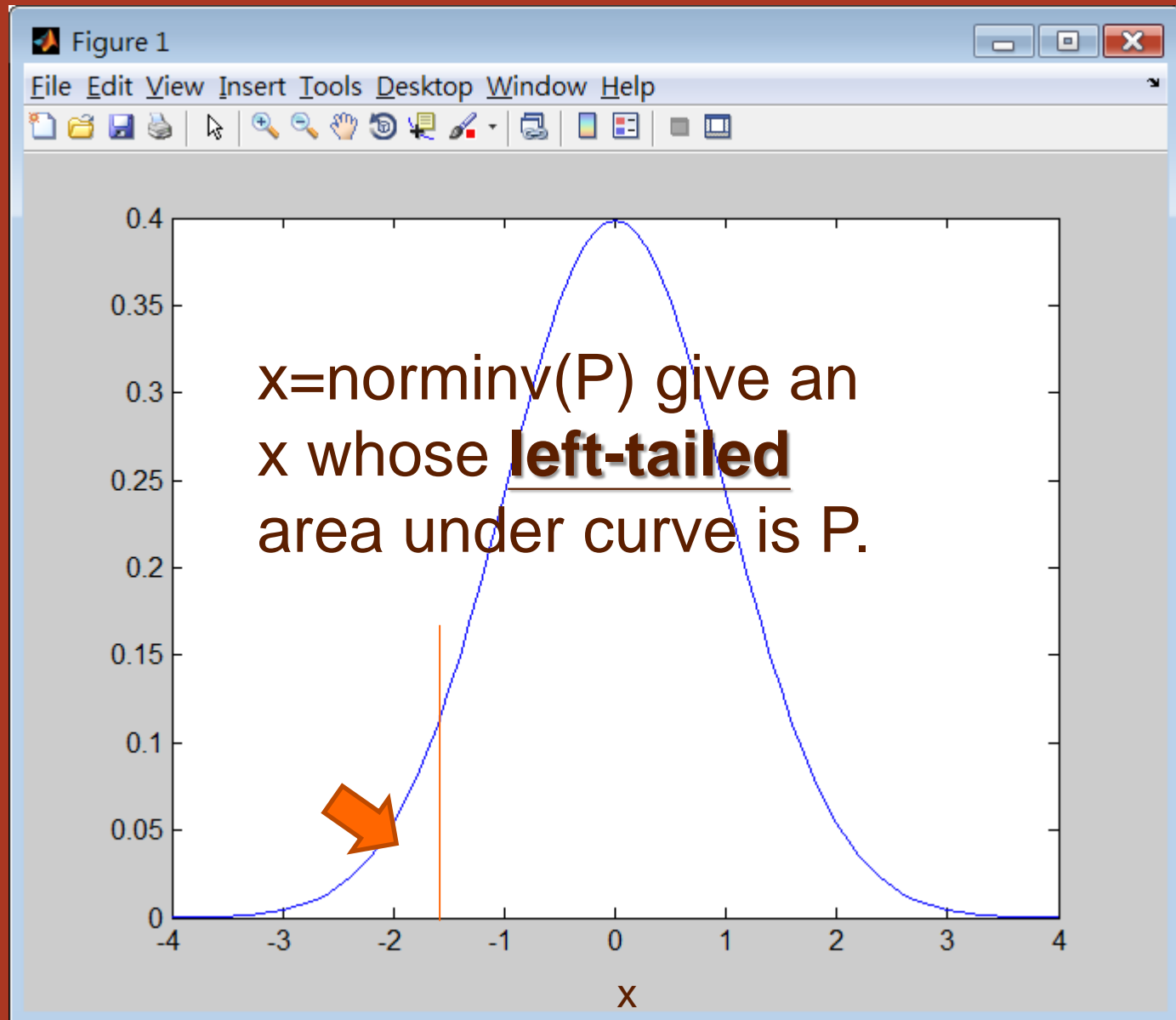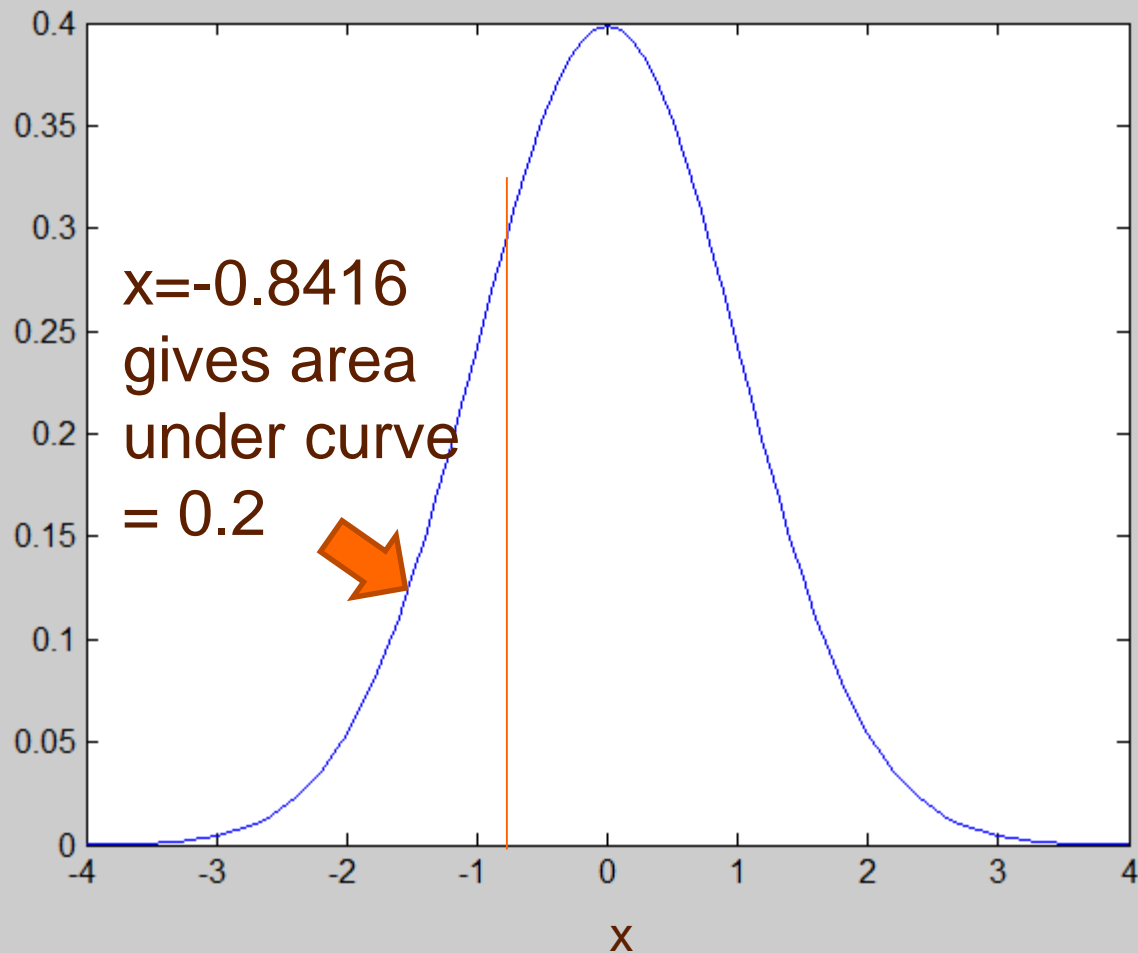
x=-0.8416 gives area under curve = 0.2

>> norminv(0.2)
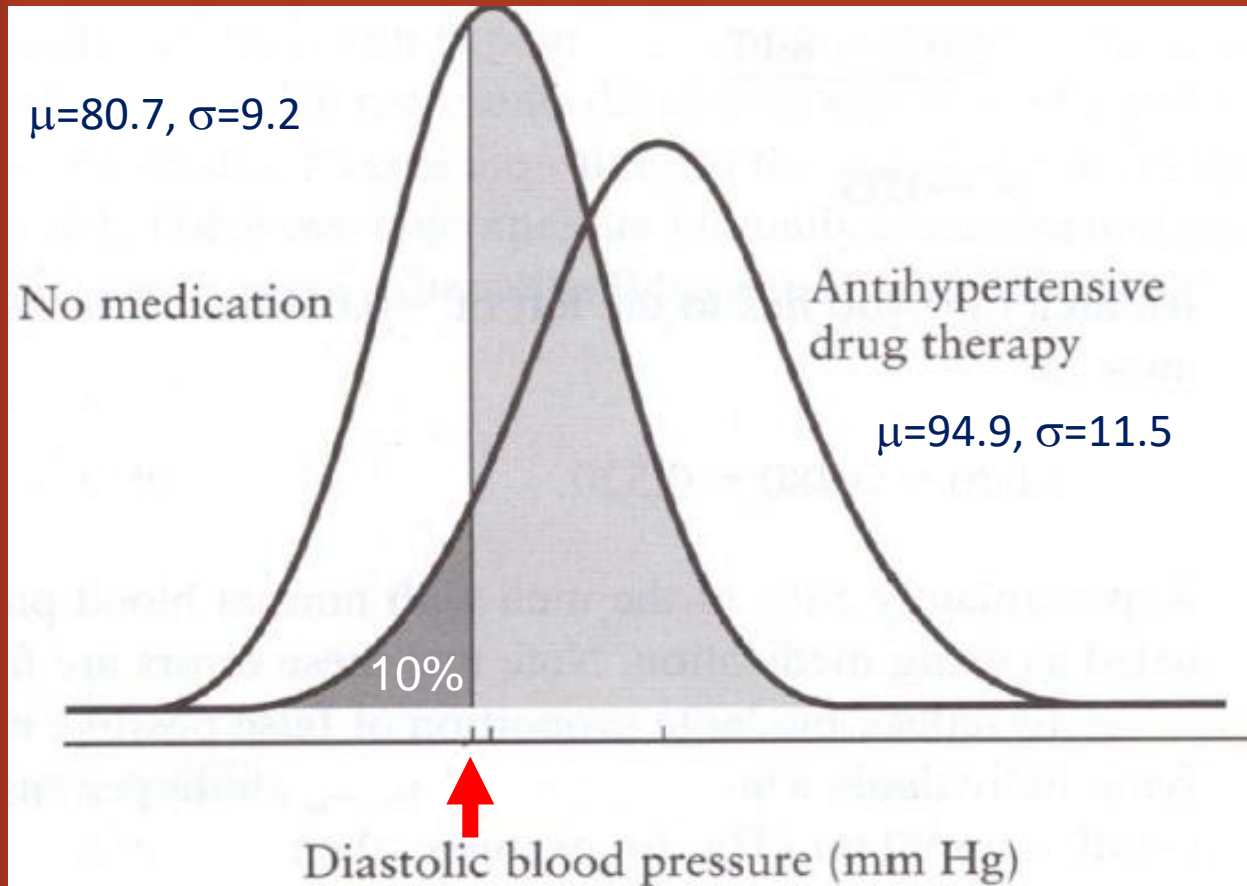ans =
    -0.8416
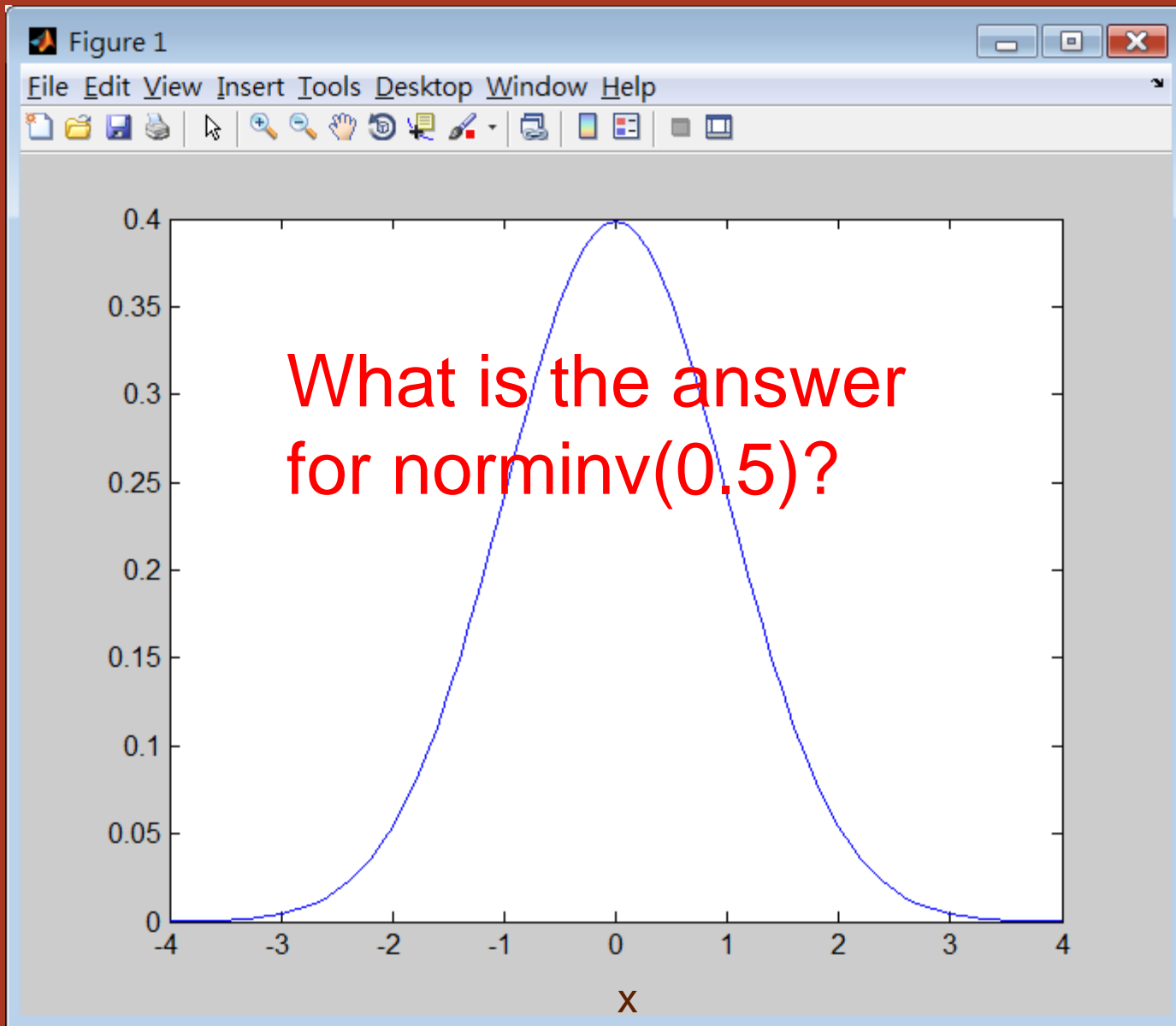
Example 5

- We have 2 normal distributions – one with people having normal blood pressure, and the other with high blood pressure and taking medication at the same time.
- The two distributions have different μ and σ (next slide).
- Our goal is to know whether one has normal blood pressure or is taking antihypertensive drugs, solely on the basis of reading his blood pressure.
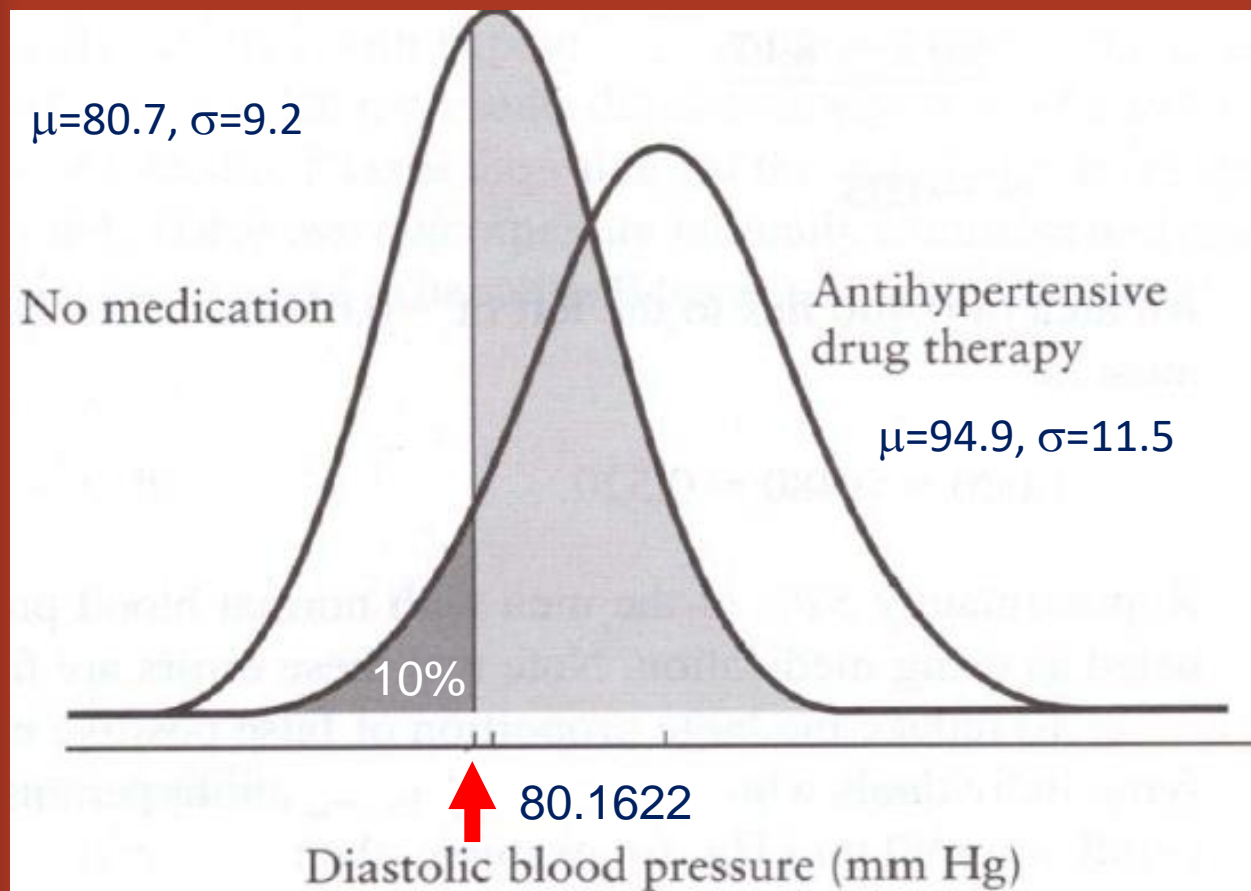
μ=80.7, σ=9.2

No medication

Antihypertensive drug therapy

μ=94.9, σ=11.5

10%

Diastolic blood pressure (mm Hg)
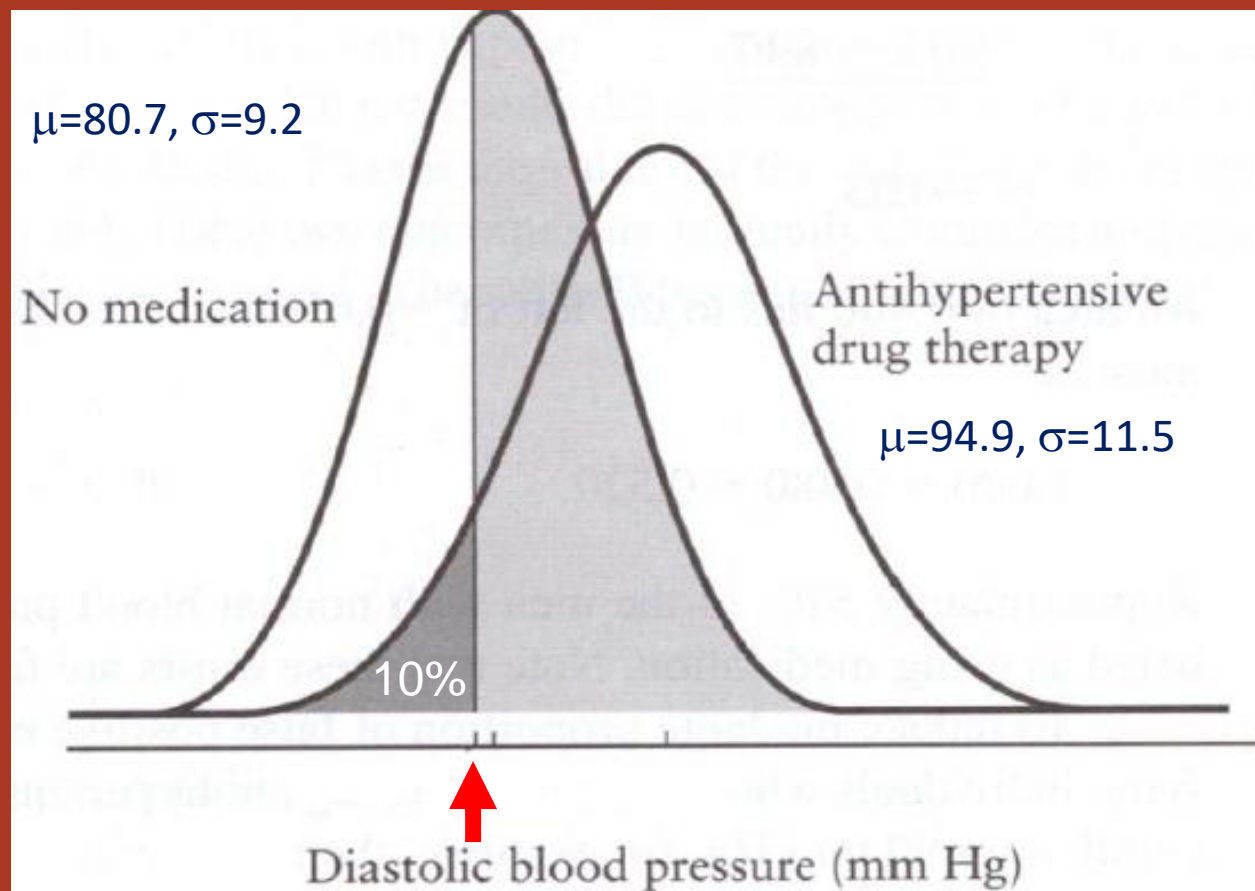
Let's locate the lower 10% of the 'medication' group and find this blood pressure reading. Below this mark, one person is **not likely** to be under medication.

μ=80.7, σ=9.2

No medication

Antihypertensive drug therapy

μ=94.9, σ=11.5

10%

Diastolic blood pressure (mm Hg)

>> **norminv(0.1, 94.9, 11.5)**
ans =   80.1622
>>

μ=80.7, σ=9.2

No medication

Antihypertensive drug therapy

μ=94.9, σ=11.5

10%

80.1622

Diastolic blood pressure (mm Hg)

Using this mark 80.1622, however, would **falsely** identify a 'big' portion of normal people as ones taking medication (light gray area). How big is this portion?
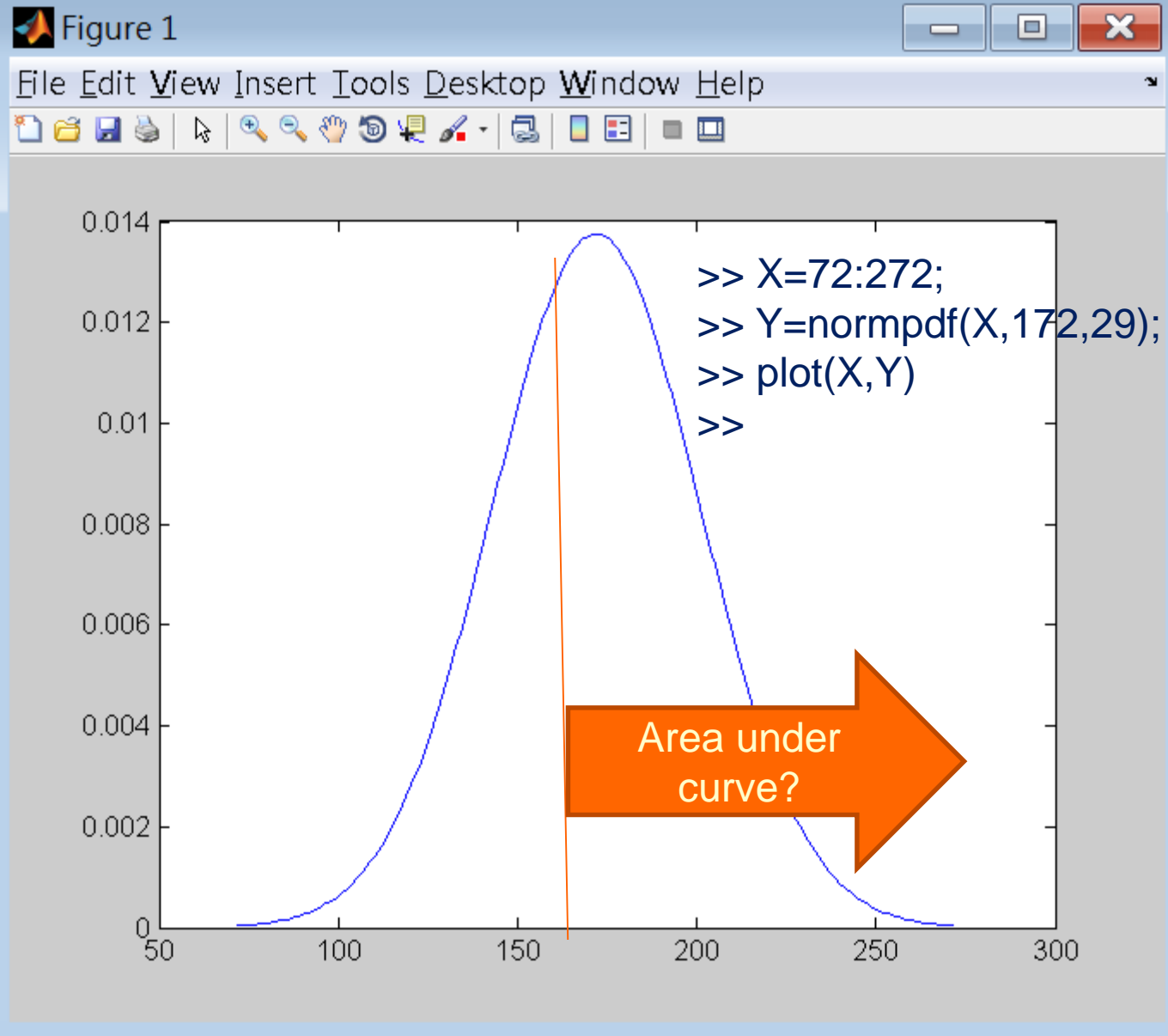
μ=80.7, σ=9.2

No medication

Antihypertensive drug therapy

μ=94.9, σ=11.5

10%

80.1622

Diastolic blood pressure (mm Hg)
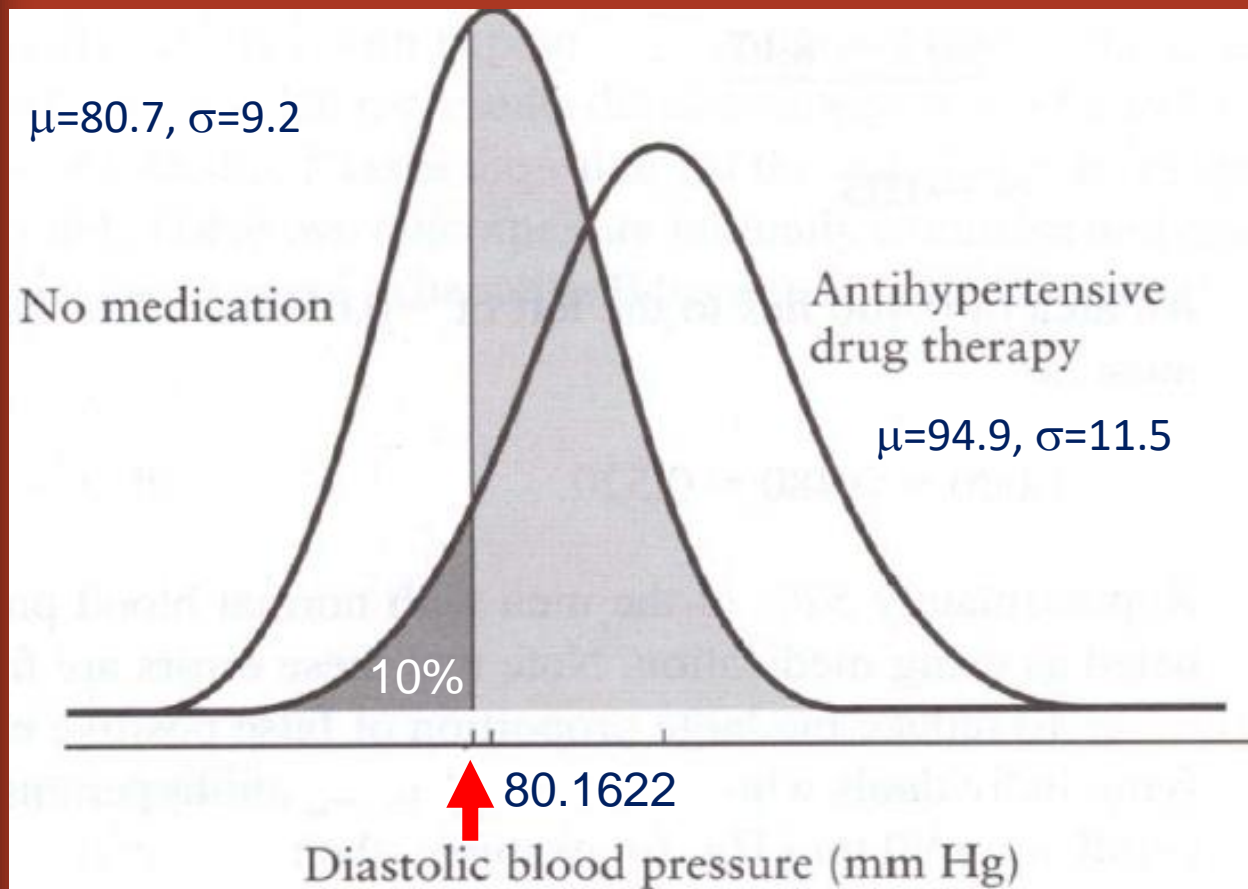
ans=80.1622
from previous slide

Area light gray:
>> 1-normcdf(ans, 80.7, 9.2)
ans = 0.5233

- Note that Example 5 was meant to illustrate normcdf() and norminv().
- It has nothing to do with CI we intended to introduced in this lecture. (See there is no sample size n involved~~~)

Example 6

- Weight limit for an elevator is 12 persons with each weighing 167 pounds.
- Men have weights that are normally distributed with a mean of 172 pounds and a standard deviation of 29 pounds.
- Q1: Probability for one man weighing over 167 pounds?
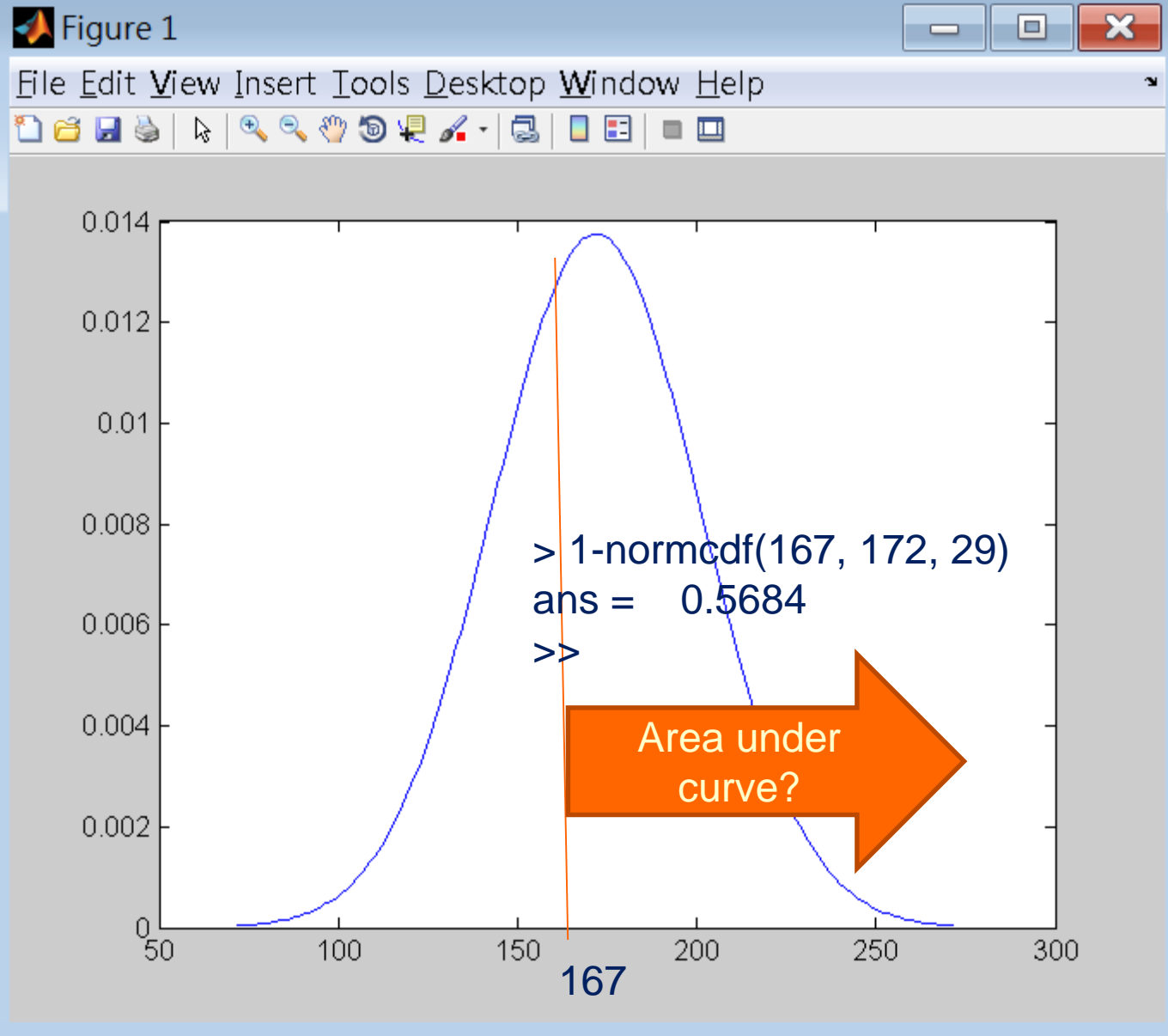- Q2: Probability for an average weight from a random sample of 12 men over 167 pounds?

Figure 1

For Q2: A normal distribution with mean = 172 and STD=29/sqrt(12)

>> 1-normcdf(167, 172, 29/sqrt(12))
ans =   0.7248
>>

Area under green curve?

167