Sampling distributions

Sampling distributions

We will repeat our main goal many times in many ways.

Sampling distributions

We will repeat our main goal many times in many ways. Here's an easy question to remember: how does one compute the "margin of error" for a poll?

Sampling distributions

We will repeat our main goal many times in many ways. Here's an easy question to remember: how does one compute the "margin of error" for a poll? How does Gallup know that 65% plus or minus 4% of Americans like chocolate chip cookies?

Sampling distributions

We will repeat our main goal many times in many ways. Here's an easy question to remember: how does one compute the "margin of error" for a poll? How does Gallup know that 65% plus or minus 4% of Americans like chocolate chip cookies? Do they really know that, anyways?

Sampling distributions

We will repeat our main goal many times in many ways. Here's an easy question to remember: how does one compute the "margin of error" for a poll? How does Gallup know that 65% plus or minus 4% of Americans like chocolate chip cookies? Do they really know that, anyways?

Remember from last time that we are trying to understand a parameter (like the true average of purchase prices for cars in the U.S.) from a statistic (the

average of say 1000 of those purchases picked at random.

average of say 1000 of those purchases picked at random.

The main conceptual turn: we think of a statistic as a random variable.

average of say 1000 of those purchases picked at random.

The main conceptual turn: we think of a statistic as a random variable. After all, choosing 1000 car purchases at random and taking (for example) the mean price is akin to rolling dice - unpredictable, but over many (thousands) of random samples we expect to see the true mean purchase price on average.

average of say 1000 of those purchases picked at random.

The main conceptual turn: we think of a statistic as a random variable. After all, choosing 1000 car purchases at random and taking (for example) the mean price is akin to rolling dice - unpredictable, but over many (thousands) of random samples we expect to see the true mean purchase price on average.

Taking this conceptual leap, we now look at all possible samples from a given collection and see what happens to the statistics.

average of say 1000 of those purchases picked at random.

The main conceptual turn: we think of a statistic as a random variable. After all, choosing 1000 car purchases at random and taking (for example) the mean price is akin to rolling dice - unpredictable, but over many (thousands) of random samples we expect to see the true mean purchase price on average.

Taking this conceptual leap, we now look at all possible samples from a given collection and see what happens to the statistics.

Definition 1. *Given a population, the sampling distribution for samples of size* n *is the probability distribution of some parameter of the samples (for example the mean) as we go through all possible samples.*

Definition 1. *Given a population, the sampling distribution for samples of size n is the probability distribution of some parameter of the samples (for example the mean) as we go through all possible samples.*

We go thoroughly through an example looking at sampling distributions and comment thoroughly as we go along.

Consider the collection of numbers

$$\mathbf{S} = \{2, 2, 5, 6, 7\}.$$

What is the mean?

Consider the collection of numbers

$$\mathbf{S} = \{2, 2, 5, 6, 7\}.$$

What is the mean? Ans: 4.4.

Consider the collection of numbers

$$\mathbf{S} = \{2, 2, 5, 6, 7\}.$$

What is the mean? Ans: 4.4.

We can consider all samples from our collection of size 1. There are 5 of them and again, the mean of the samples is again 4.4. The standard deviation is 2.302.

We can now consider all samples from $\mathbf{S}$ of size 2, and take the mean of each sample.

We can now consider all samples from $\mathbf{S}$ of size 2, and take the mean of each sample. Our samples are

$$\{\{2,2\},\{2,5\},\{2,6\},\{2,7\},\{2,5\},\{2,6\},$$
$$\{2,7\},\{5,6\},\{5,7\},\{6,7\}\}.$$

We can now consider all samples from $\mathbf{S}$ of size 2, and take the mean of each sample. Our samples are

$$\{\{2,2\},\{2,5\},\{2,6\},\{2,7\},\{2,5\},\{2,6\},$$
$$\{2,7\},\{5,6\},\{5,7\},\{6,7\}\}.$$

Our means are

$$\{2, 3.5, 4, 4.5, 3.5, 4, 4.5, 5.5, 6, 6.5\}$$

We can now consider all samples from $\mathbf{S}$ of size 2, and take the mean of each sample. Our samples are

$$\{\{2,2\},\{2,5\},\{2,6\},\{2,7\},\{2,5\},\{2,6\},\{2,7\},\{5,6\},\{5,7\},\{6,7\}\}.$$

Our means are

$$\{2, 3.5, 4, 4.5, 3.5, 4, 4.5, 5.5, 6, 6.5\}$$

Notice that the spread is smaller, even though there are more numbers. The mean is still 4.4, and the standard deviation is now 1.329.

We can now consider all samples from $\mathbf{S}$ of size 2, and take the mean of each sample. Our samples are

$$\{\{2,2\},\{2,5\},\{2,6\},\{2,7\},\{2,5\},\{2,6\},\{2,7\},\{5,6\},\{5,7\},\{6,7\}\}.$$

Our means are

$$\{2, 3.5, 4, 4.5, 3.5, 4, 4.5, 5.5, 6, 6.5\}$$

Notice that the spread is smaller, even though there are more numbers. The mean is still 4.4, and the standard deviation is now 1.329.

The probability distribution of these numbers is an example of sampling distribution.

The probability distribution of these numbers is an example of sampling distribution. In this case it is for the mean of a sample of 2 from the set $\mathbf{S}$.

The probability distribution of these numbers is an example of sampling distribution. In this case it is for the mean of a sample of 2 from the set $\mathbf{S}$. There are 7 possible outcomes which *don't* have equal probabilities.

The probability distribution of these numbers is an example of sampling distribution. In this case it is for the mean of a sample of 2 from the set $\mathbf{S}$. There are 7 possible outcomes which *don't* have equal probabilities.

X	2	3.5	4	4.5	5.5	6	6.5
Probability	.1	.2	.2	.2	.1	.1	.1

The probability distribution of these numbers is an example of sampling distribution. In this case it is for the mean of a sample of 2 from the set $\mathbf{S}$. There are 7 possible outcomes which *don't* have equal probabilities.

X	2	3.5	4	4.5	5.5	6	6.5
Probability	.1	.2	.2	.2	.1	.1	.1

Now consider all samples of size 3, and their means.

Now consider all samples of size 3, and their means. There are again 10 such samples (you might recall this from our digression on binomial coefficients), which we won't list.

Now consider all samples of size 3, and their means. There are again 10 such samples (you might recall this from our digression on binomial coefficients), which we won't list. The means for these 10 samples are

$$\{3, 3.333, 3.667, 4.333, 4.667, 4.333, 4.667, 5, 5, 6\}$$

Now consider all samples of size 3, and their means. There are again 10 such samples (you might recall this from our digression on binomial coefficients), which we won't list. The means for these 10 samples are

$$\{3, 3.333, 3.667, 4.333, 4.667, 4.333, 4.667, 5, 5, 6\}$$

The standard deviation is now only .886.

Now consider all samples of size 3, and their means. There are again 10 such samples (you might recall this from our digression on binomial coefficients), which we won't list. The means for these 10 samples are

$$\{3, 3.333, 3.667, 4.333, 4.667, 4.333, 4.667, 5, 5, 6\}$$

The standard deviation is now only .886.

Our sampling distribution for samples of 3 out of our original group S is

X	3	3.333	3.667	4.333	4.667	5	6
Probability	.1	.1	.1	.2	.2	.2	.1

X	3	3.333	3.667	4.333	4.667	5	6
Probability	.1	.1	.1	.2	.2	.2	.1

Example 2. *Calculate the sampling distribution for the mean of samples of three out of the data set* $\{-1, 0, 0, 2, 3, 6\}$.

X	3	3.333	3.667	4.333	4.667	5	6
Probability	.1	.1	.1	.2	.2	.2	.1

Example 2. *Calculate the sampling distribution for the mean of samples of three out of the data set* $\{-1, 0, 0, 2, 3, 6\}$.

Facts about sampling distributions

We see in these examples that the sampling distribution of size n always has the same mean as the original list

Facts about sampling distributions

We see in these examples that the sampling distribution of size n always has the same mean as the original list and has a standard deviation which is decreasing as the size gets larger.

Facts about sampling distributions

We see in these examples that the sampling distribution of size n always has the same mean as the original list and has a standard deviation which is decreasing as the size gets larger.

Theorem 3. *Let X be a population with mean μ and standard deviation σ.*

Facts about sampling distributions

We see in these examples that the sampling distribution of size n always has the same mean as the original list and has a standard deviation which is decreasing as the size gets larger.

Theorem 3. *Let X be a population with mean μ and standard deviation σ.*

Consider the sampling distribution of means of samples of size n from our population.

Facts about sampling distributions

We see in these examples that the sampling distribution of size n always has the same mean as the original list and has a standard deviation which is decreasing as the size gets larger.

Theorem 3. *Let X be a population with mean μ and standard deviation σ.*

Consider the sampling distribution of means of samples of size n from our population.

- *The mean of the sampling distribution is again* μ.

- *The mean of the sampling distribution is again* μ.
- *The standard distribution of the sampling distribution is* $\sigma/\sqrt{n}$.

- *The mean of the sampling distribution is again* μ.
- *The standard distribution of the sampling distribution is* $\sigma/\sqrt{n}$.
- *If our population is* $N(\mu, \sigma)$ *then our sampling distribution is* $N(\mu, \sigma/\sqrt{n})$.

- *The mean of the sampling distribution is again μ.*
- *The standard distribution of the sampling distribution is $\sigma/\sqrt{n}$.*
- *If our population is $N(\mu, \sigma)$ then our sampling distribution is $N(\mu, \sigma/\sqrt{n})$.*

These features of sampling distributions, which are true in general, will be critical in making statistical inferences.

- *The mean of the sampling distribution is again μ.*
- *The standard distribution of the sampling distribution is $\sigma/\sqrt{n}$.*
- *If our population is $N(\mu, \sigma)$ then our sampling distribution is $N(\mu, \sigma/\sqrt{n})$.*

These features of sampling distributions, which are true in general, will be critical in making statistical inferences.

The Central Limit Theorem

The Central Limit Theorem

Theorem 4. *The sampling distribution of means of random samples of size n from a population with mean μ and standard deviation σ is approximately*

$$N(\mu, \sigma/\sqrt{n})$$

when n is large.

The Central Limit Theorem

Theorem 4. *The sampling distribution of means of random samples of size n from a population with mean μ and standard deviation σ is approximately*

$$N(\mu, \sigma/\sqrt{n})$$

when n is large.

This theorem is true no matter what the original distribution of our population is!

The Central Limit Theorem

Theorem 4. *The sampling distribution of means of random samples of size n from a population with mean μ and standard deviation σ is approximately*

$$N(\mu, \sigma/\sqrt{n})$$

when n is large.

This theorem is true no matter what the original distribution of our population is! (unlike the previous theorem)

This theorem is key to the kingdom of statistics.

This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation

This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation - getting smaller.

This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation - getting smaller. What does that tell us?

This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation - getting smaller. What does that tell us? Knowing the deviation of the sampling distribution and the fact that it is approximately normal, we know how likely it is that a sample is within that deviation (or some multiple) from the mean.

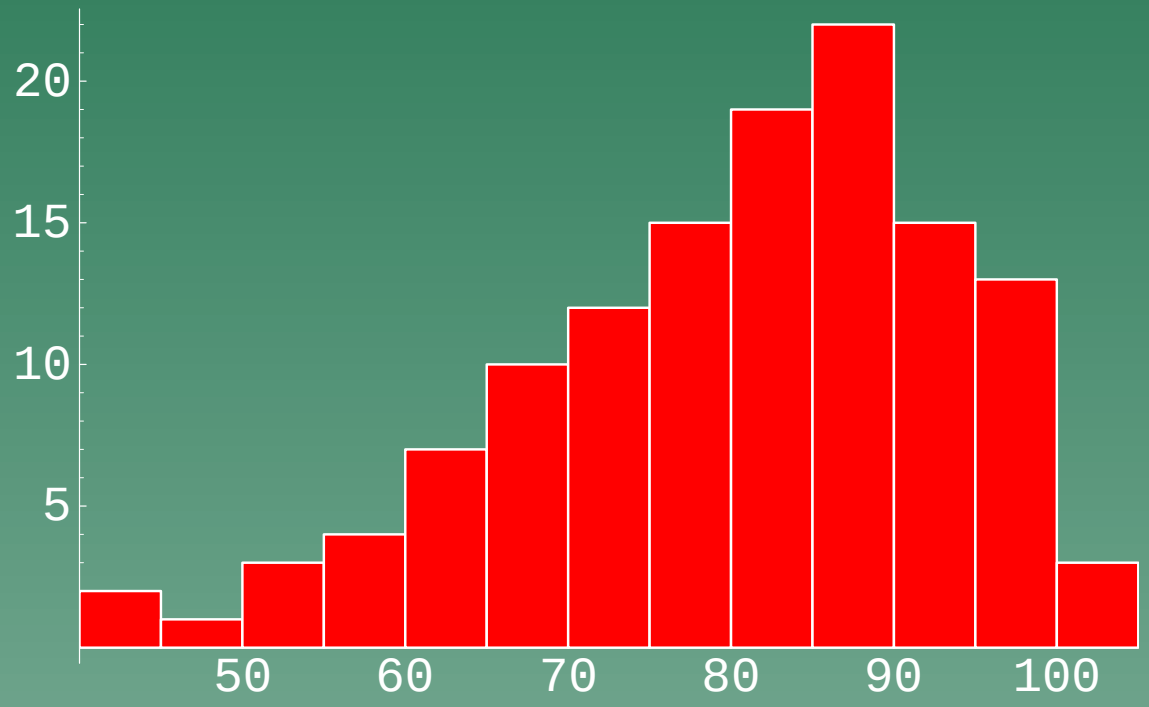
This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation - getting smaller. What does that tell us? Knowing the deviation of the sampling distribution and the fact that it is approximately normal, we know how likely it is that a sample is within that deviation (or some multiple) from the mean. So we can understand our basic question: how far our sample mean probably is from the real mean.

This theorem is key to the kingdom of statistics. Think of what is happening to the standard deviation - getting smaller. What does that tell us? Knowing the deviation of the sampling distribution and the fact that it is approximately normal, we know how likely it is that a sample is within that deviation (or some multiple) from the mean. So we can understand our basic question: how far our sample mean probably is from the real mean.

Before we do precisely these kinds of computations, let's see the Central Limit Theorem in action.

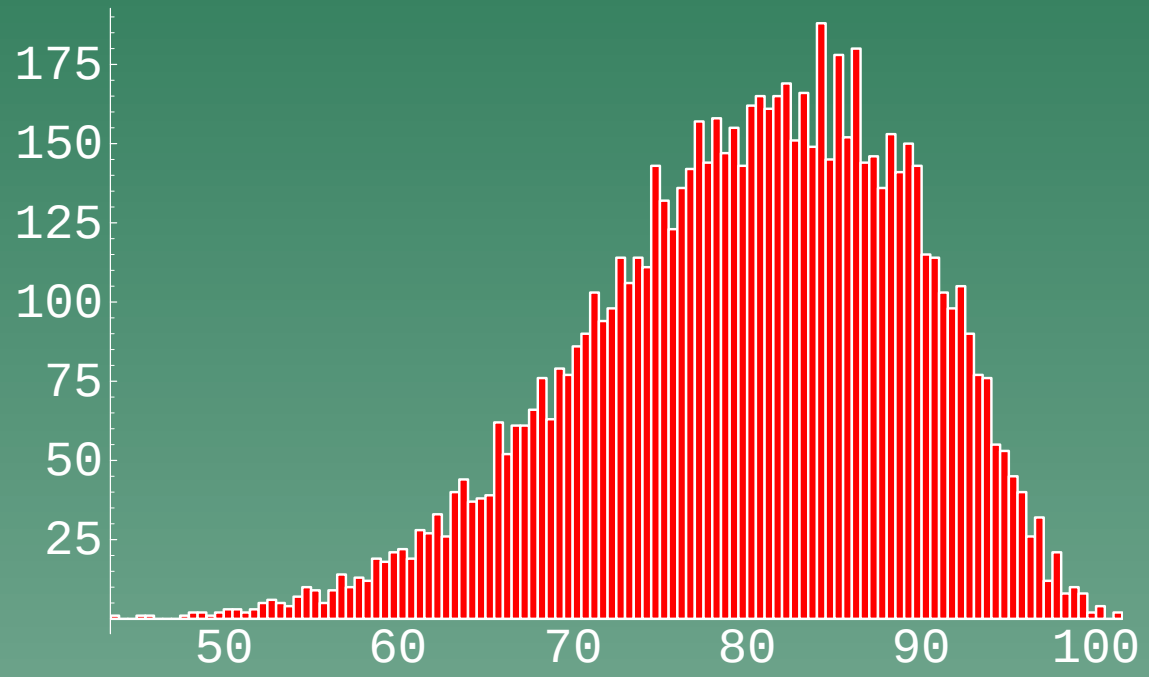
Example 5. *We look at the grade distribution for an exam.*

Example 5. *We look at the grade distribution for an exam.*

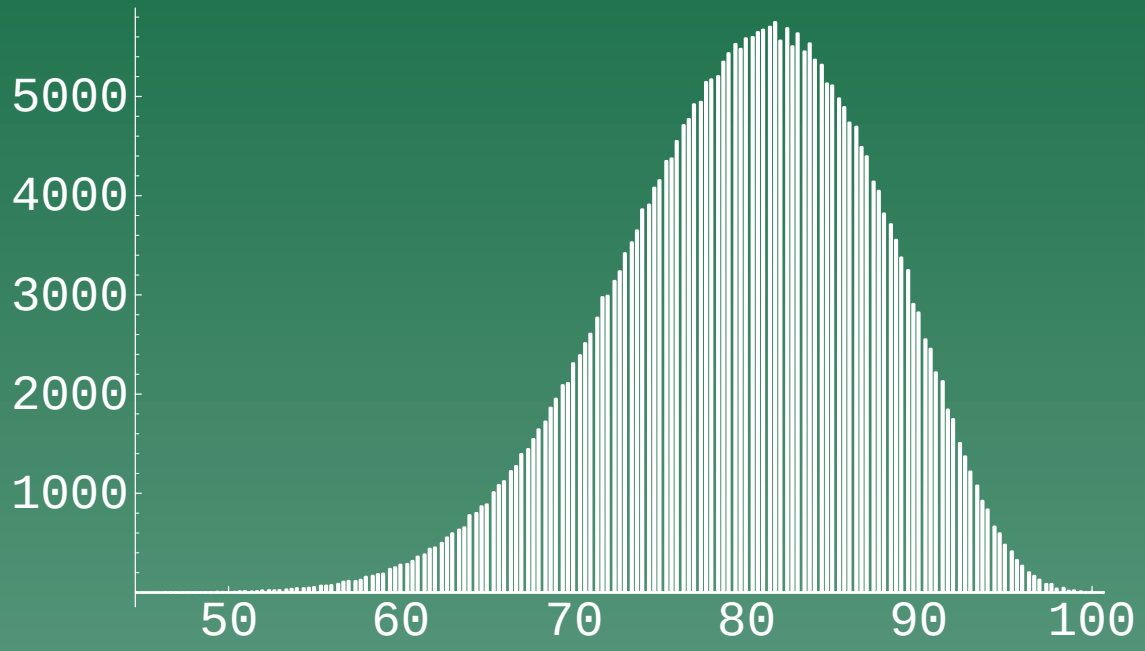


Now we take means of all samples of size 2, and look at the histogram of those numbers. (There are 7875 such samples.)

Now we take means of all samples of size 2, and look at the histogram of those numbers. (There are 7875 such samples.)



Now samples of size 3. (325500)



Example 6. *Suppose that the average price of a new car purchase is $24145 with a standard deviation of $3615.*

Example 6. *Suppose that the average price of a new car purchase is $24145 with a standard deviation of $3615. Suppose you take a survey of 1000 car purchases. What is the probability that the average over your survey is over $25000?*

Example: Process Control

Example: Process Control

The Central Limit Theorem has many applications, since sampling can be useful well beyond the realms of surveys and opinion polls.

Example: Process Control

The Central Limit Theorem has many applications, since sampling can be useful well beyond the realms of surveys and opinion polls.

- Imagine a manufacturing process for, say, ball bearings. The bearings are supposed to be 10 mm in diameter. In fact, when the manufacturing process is working correctly, they are distributed normally $N(10, .7)$.

- We can't check every bearing as it is too time consuming. Every hour we take a sample of 10 bearings, and take the mean diameter. The sample distribution of the means, $\overline{x}$ should be $N(10, .7/\sqrt{10}) = N(10, .221)$.

- We can't check every bearing as it is too time consuming. Every hour we take a sample of 10 bearings, and take the mean diameter. The sample distribution of the means, $\overline{x}$ should be $N(10, .7/\sqrt{10}) = N(10, .221)$.

- This means (by the 68-95-99.7 rule) 99.7% of the means will occur

$$10 - 3(.221) < \overline{x} < 10 + 3(.221)$$

$$9.337 < \overline{x} < 10.663$$

- We can't check every bearing as it is too time consuming. Every hour we take a sample of 10 bearings, and take the mean diameter. The sample distribution of the means, $\overline{x}$ should be $N(10, .7/\sqrt{10}) = N(10, .221)$.

- This means (by the 68-95-99.7 rule) 99.7% of the means will occur

$$10 - 3(.221) < \overline{x} < 10 + 3(.221)$$

$$9.337 < \overline{x} < 10.663$$

We are alarmed if we see any $\overline{x}$ outside of these limits, and suspect that our manufacturing process has been disturbed.

We are alarmed if we see any $\overline{x}$ outside of these limits, and suspect that our manufacturing process has been disturbed.

- We keep track with a $\overline{x}$ control chart. This is a graph, with a mark each hour for the value of $\overline{x}$ for that hour.

We are alarmed if we see any $\overline{x}$ outside of these limits, and suspect that our manufacturing process has been disturbed.

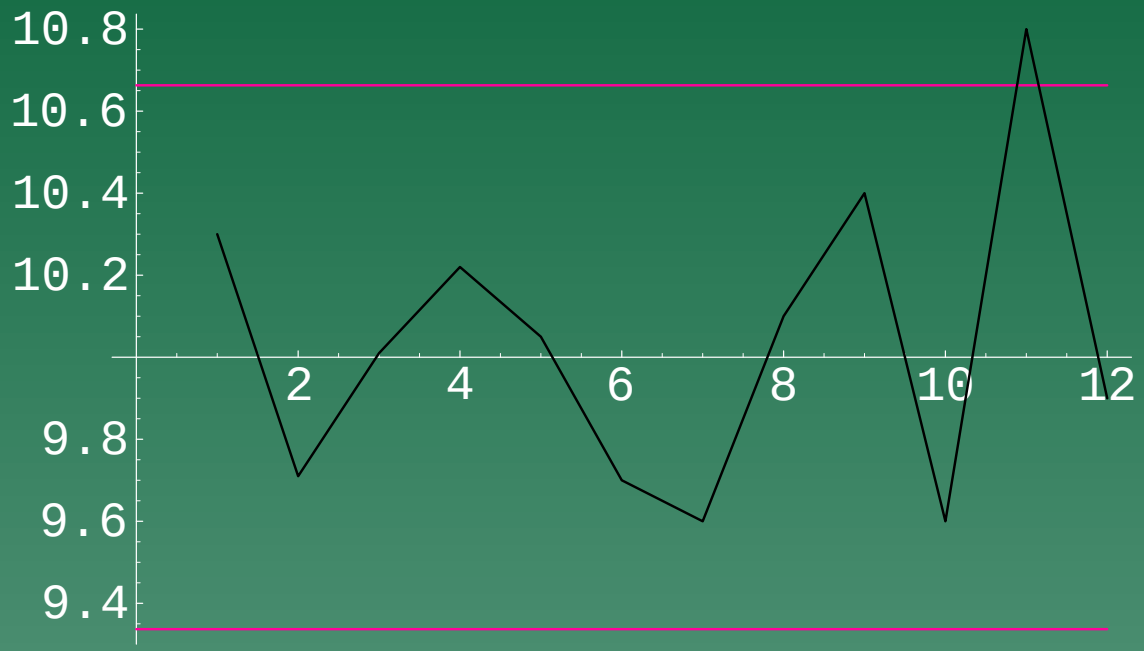
- We keep track with a $\overline{x}$ control chart. This is a graph, with a mark each hour for the value of $\overline{x}$ for that hour.

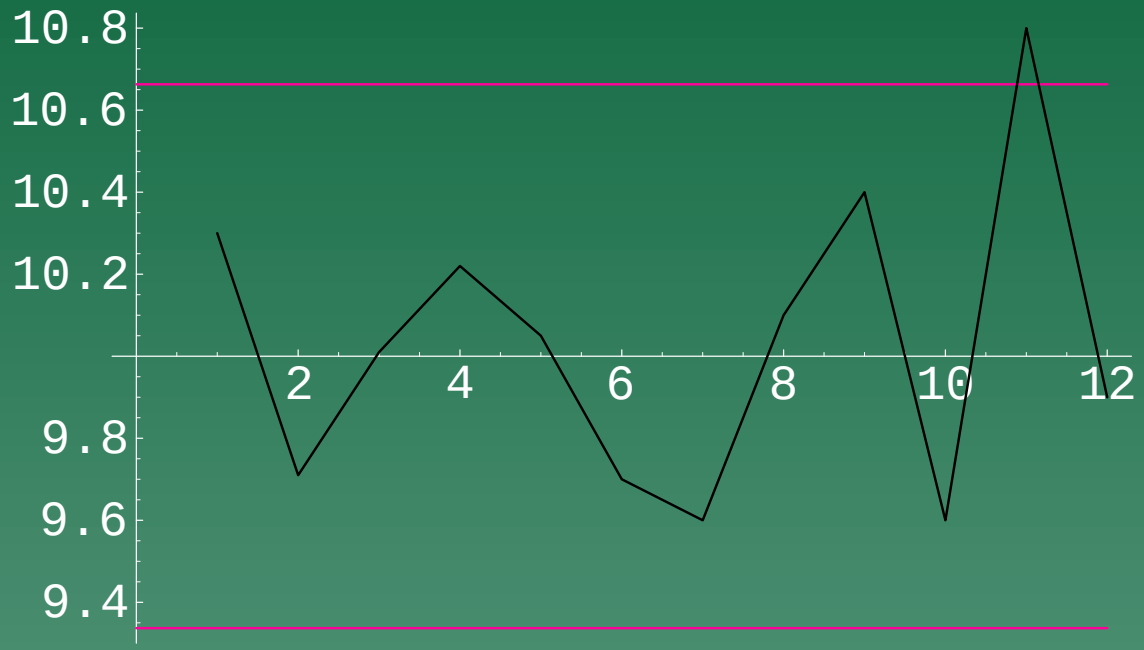
  The graph includes an upper control line 3 standard deviations above the mean, and a lower control lines, 3 standard deviations below the mean.

We are alarmed if we see any $\overline{x}$ outside of these limits, and suspect that our manufacturing process has been disturbed.

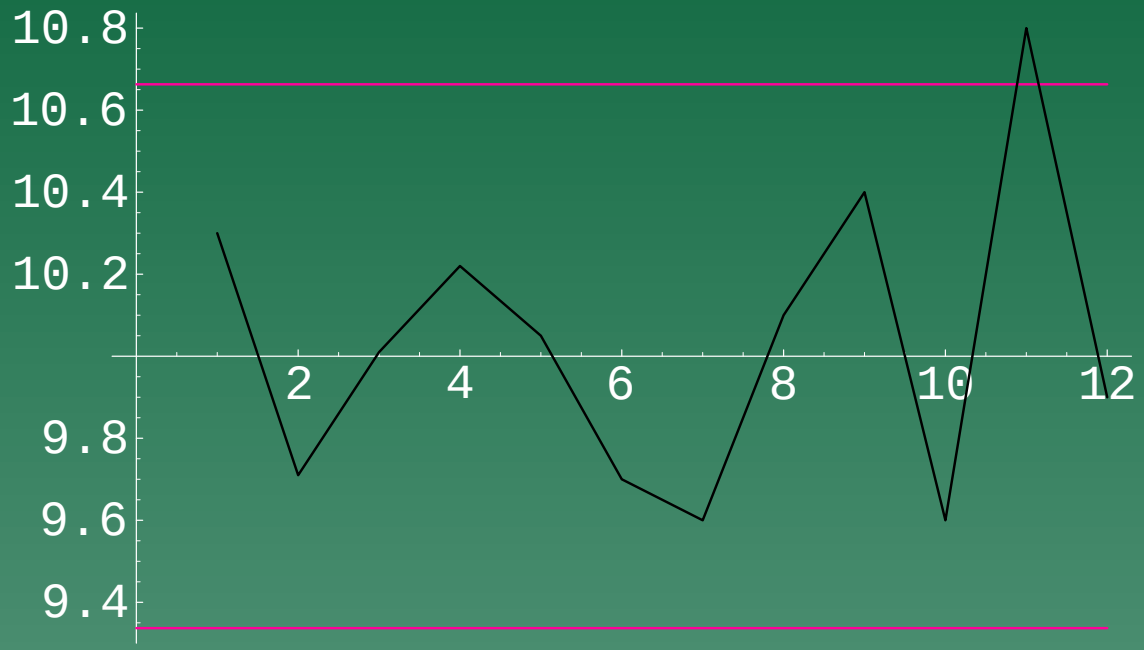
- We keep track with a $\overline{x}$ control chart. This is a graph, with a mark each hour for the value of $\overline{x}$ for that hour.

  The graph includes an upper control line 3 standard deviations above the mean, and a lower control lines, 3 standard deviations below the mean.


10.8
10.6
10.4
10.2
2
4
6
8
10
12
9.8
9.6
9.4



Notice the entry above the upper control line. Since 99.7% of the entries should be between the control lines, we should only get entries outside of that range about 3/1000 times.



Notice the entry above the upper control line. Since 99.7% of the entries should be between the control lines, we should only get entries outside of that range about 3/1000 times.

So an entry above the upper control line should be a rare event and should mean that we check our production line to see if problems have developed.

So an entry above the upper control line should be a rare event and should mean that we check our production line to see if problems have developed.