Using the Central Limit Theorem

Using the Central Limit Theorem

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

Using the Central Limit Theorem

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?

Using the Central Limit Theorem

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?

Using the Central Limit Theorem

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.

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

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.

The main - the only, really - theoretical basis we'll see for statistical inference is the Central Limit Theorem.

The main - the only, really - theoretical basis we'll see for statistical inference is the Central Limit Theorem.

Theorem 1. *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.*

Example 2. *Suppose that the average price of a new car purchase is \$24145 with a standard deviation of \$3615.*

Example 2. *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 3. *If SAT scores are distributed normally according to* $N(1630, 100)$, *what is the chance of six* randomly *sampled students having an average score above 1800?*

Example 3. *If SAT scores are distributed normally according to* $N(1630, 100)$*, what is the chance of six* randomly *sampled students having an average score above 1800? (If you keep on "sampling" students and finding a larger average than 1800, what should you deduce?)*

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.

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

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

We can't check every bearing.

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

We can't check every bearing. 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.

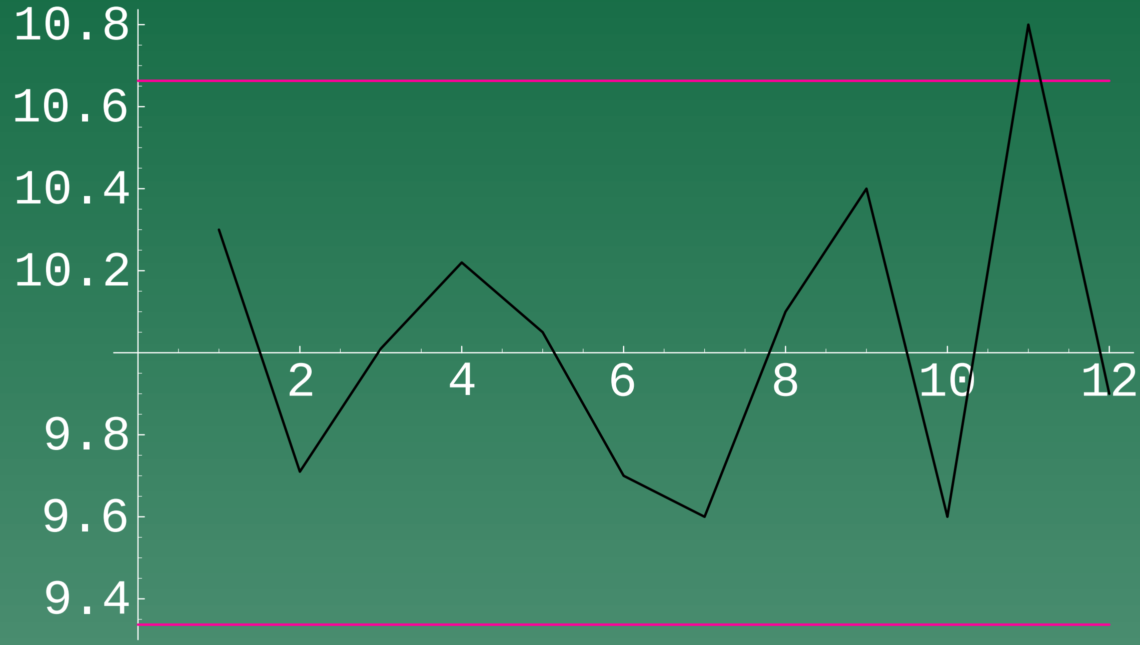
We keep track with a $\overline{x}$ control chart. This is a graph, with a mark each hour for the value of $\overline{x}$ for that hour.

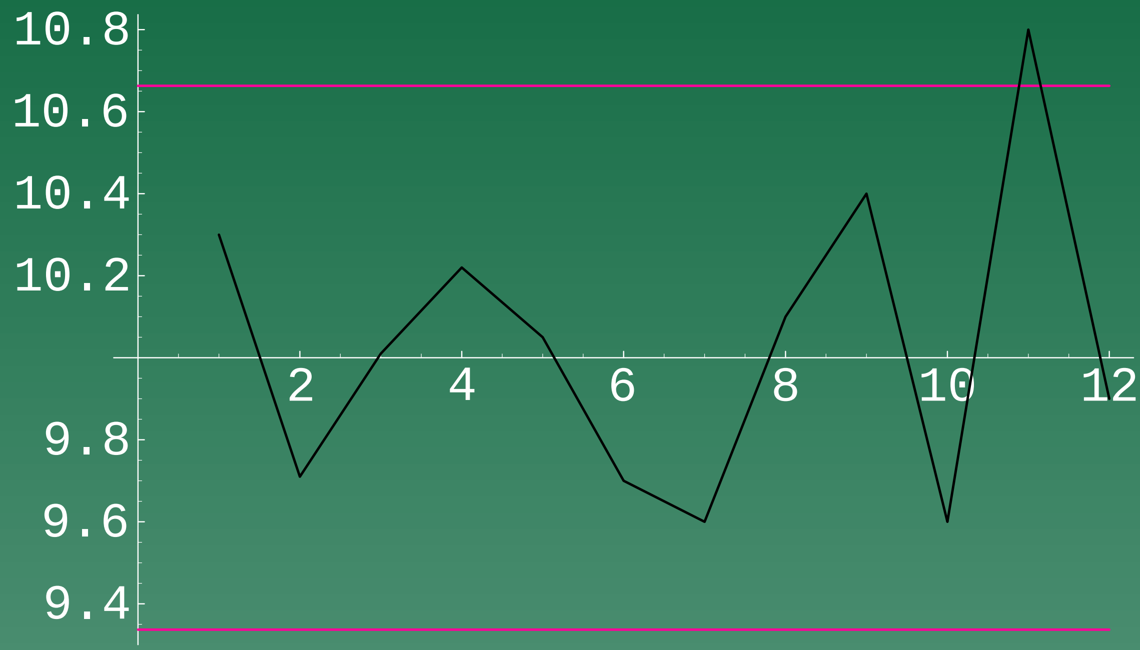
The graph includes an upper control line 3 standard deviations above the mean, and a lower control lines, 3 standard deviations below the mean.

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
9.8
9.6
9.4
2
4
6
8
10
12



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.

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.

Confidence intervals

Confidence intervals

Confidence intervals are one of the most basic parts of the analysis we do for the rest of the term.

Confidence intervals

Confidence intervals are one of the most basic parts of the analysis we do for the rest of the term.

Definition 4. *A confidence interval is a statement of the form "There is a C% chance that the parameter we are trying to measure is between* $X - D$ *and* $X + D$*."*

Confidence intervals

Confidence intervals are one of the most basic parts of the analysis we do for the rest of the term.

Definition 4. *A confidence interval is a statement of the form "There is a C% chance that the parameter we are trying to measure is between* $X - D$ *and* $X + D$*." The quantity* D *is called the margin of error.*

Confidence intervals

Confidence intervals are one of the most basic parts of the analysis we do for the rest of the term.

Definition 4. *A confidence interval is a statement of the form "There is a C% chance that the parameter we are trying to measure is between* $X - D$ *and* $X + D$*." The quantity* D *is called the margin of error. The value* X *will most often be a statistical measure of the parameter through some survey.*

For example, “There is a 95% chance that the between 30 and 34% of registered voters approve of the job President Bush is doing.”

For example, “There is a 95% chance that the between 30 and 34% of registered voters approve of the job President Bush is doing.” This is pretty close to the standard statement “32% of registered voters surveyed approve of the job President Bush is doing, with a margin of error of $\pm 2\%$.

For example, "There is a 95% chance that the between 30 and 34% of registered voters approve of the job President Bush is doing." This is pretty close to the standard statement "32% of registered voters surveyed approve of the job President Bush is doing, with a margin of error of $\pm 2\%$. (The standard statement leaves out the "95% chance bit" - we will see that they are being a bit misleading.)

We would like to construct confidence intervals just from having survey data.

We would like to construct confidence intervals just from having survey data. But what we do as a first step is to construct them *assuming we already know the standard deviation of the distribution from which they are sampled* by using the Central Limit Theorem.

We would like to construct confidence intervals just from having survey data. But what we do as a first step is to construct them *assuming we already know the standard deviation of the distribution from which they are sampled* by using the Central Limit Theorem.

Example 5. *Suppose we know that the standard deviation for new car purchases is $3000*

We would like to construct confidence intervals just from having survey data. But what we do as a first step is to construct them *assuming we already know the standard deviation of the distribution from which they are sampled* by using the Central Limit Theorem.

Example 5. *Suppose we know that the standard deviation for new car purchases is \$3000 and we run a survey of 700 purchases which finds a mean purchase price of 23,142. Find a confidence interval for the actual mean with a 95% certainty.*

We would like to construct confidence intervals just from having survey data. But what we do as a first step is to construct them *assuming we already know the standard deviation of the distribution from which they are sampled* by using the Central Limit Theorem.

Example 5. *Suppose we know that the standard deviation for new car purchases is \$3000 and we run a survey of 700 purchases which finds a mean purchase price of 23,142. Find a confidence interval for the actual mean with a 95% certainty. or with 99.7% certainty.*

We would like to construct confidence intervals just from having survey data. But what we do as a first step is to construct them *assuming we already know the standard deviation of the distribution from which they are sampled* by using the Central Limit Theorem.

Example 5. *Suppose we know that the standard deviation for new car purchases is $3000 and we run a survey of 700 purchases which finds a mean purchase price of 23,142. Find a confidence interval for the actual mean with a 95% certainty. or with 99.7% certainty.*

Our method is to use the Central Limit Theorem to find the deviation for the sample distribution;

Our method is to use the Central Limit Theorem to find the deviation for the sample distribution; then just note that if the sample mean is within two deviations of the true mean then the true mean is within two deviations of the sample mean (duh) and this happens 95% of the time;

Our method is to use the Central Limit Theorem to find the deviation for the sample distribution; then just note that if the sample mean is within two deviations of the true mean then the true mean is within two deviations of the sample mean (duh) and this happens 95% of the time; so we the confidence interval will be within two (adjusted) deviations of the survey mean.

Our method is to use the Central Limit Theorem to find the deviation for the sample distribution; then just note that if the sample mean is within two deviations of the true mean then the true mean is within two deviations of the sample mean (duh) and this happens 95% of the time; so we the confidence interval will be within two (adjusted) deviations of the survey mean.

Example 6. *We know that the standard deviation for heights of women over the entire U.S. cannot be more than* 5 *inches. Suppose that we find a random sample of 400 women which has an average height of* 63.5 *inches.*

Establish a confidence interval for the true average height of women, with a confidence of 90%.

Establish a confidence interval for the true average height of women, with a confidence of 90%.

It is remarkable how such a small sample can give a pretty small margin of error with 90% certainty.

Establish a confidence interval for the true average height of women, with a confidence of 90%.

It is remarkable how such a small sample can give a pretty small margin of error with 90% certainty. (But we will come back to talk about the many ways in which "measurements with 90% certainty" can be completely wrong.

Establish a confidence interval for the true average height of women, with a confidence of 90%.

It is remarkable how such a small sample can give a pretty small margin of error with 90% certainty. (But we will come back to talk about the many ways in which "measurements with 90% certainty" can be completely wrong.