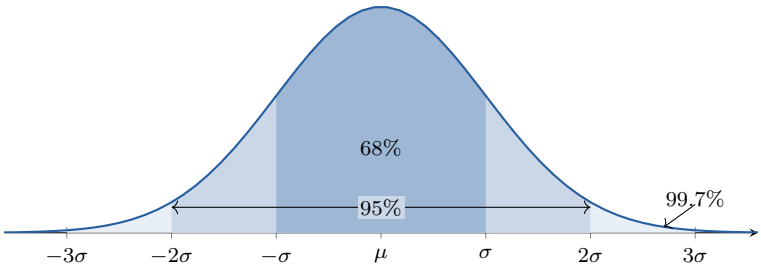
Data Analytics Workshop, JHU Pre-College Program
Day 7: Visualizing Results (optional notes)

TL;DR Today gives you the last tools for your project: how to show your answers in a picture, and how to tell from the picture whether the answer is significant.

(68-95-99.7 rule)
For normally distributed data with mean μ and standard deviation σ, about 68% of the data lies within σ of the mean, 95% within 2σ, and 99.7% within 3σ.

- The **central limit theorem**: the average of many independent samples is approximately normally distributed, with standard deviation equal to the SEM, $\frac{\sigma}{\sqrt{N}}$.
- So about 95% of the time, the mean of our data lands within $2 \cdot$SEM (more precisely $1.96 \cdot$SEM) of the true mean. This is why we draw **error bars** at $\bar{x} \pm 2 \cdot \text{SEM}$.
- To compare groups (categorical data), use a **bar chart with error bars**. If the error bars of two groups do not overlap, we have evidence the groups truly differ.
- To see how one numerical variable is spread out, use a **histogram**. To compare two numerical variables, use a **scatter plot**, and add a **line of best fit**. Its R^2 value (between 0 and 1) measures how well the line fits.



The 68-95-99.7 rule

1 The Normal Distribution and Error Bars

For your project you will make visuals that answer the questions you asked on Day 5. For comparing groups, such as average salary by race, or plant height with and without a fertilizer, a bar chart with error bars works really well. Each bar is the mean of a group, and each error bar extends a multiple of the standard error of the mean above and below it. The bars on their own are not enough: with enough variability and randomness, two means can differ just by luck. The error bars are what let us judge significance. Let's see why.

Theorem 1.1 (68-95-99.7 rule)

For normally distributed data with mean μ and standard deviation σ, about

- 68% *of the data lies within* σ *of* μ,
- 95% *lies within* 2σ *of* μ,
- 99.7% *lies within* 3σ *of* μ.

Equivalently, if the total area under the bell curve is 100%, then 68% of the area lies between $\mu - \sigma$ and $\mu + \sigma$, as on the first page. Notice how quickly the tails die off: three standard deviations account for nearly everything.

But our data is not always normal. Heights might be, but dice, coin flips, and salaries usually are not. Here is the amazing part.

Theorem 1.2 (Central limit theorem, informally)

If we take the average of N independent samples from the same distribution, then for large N that average is approximately normally distributed, no matter what the original distribution looked like.

You can see this below. On the left are 2000 random numbers between 0 and 1, every value equally likely, so the histogram is flat (uniform). On the right, each data point is instead the *average* of 10 such random numbers. Same ingredients, but now a bell curve! This is the same thing you saw with the Galton board in Discussion 3 and with binomial distributions on Day 3.

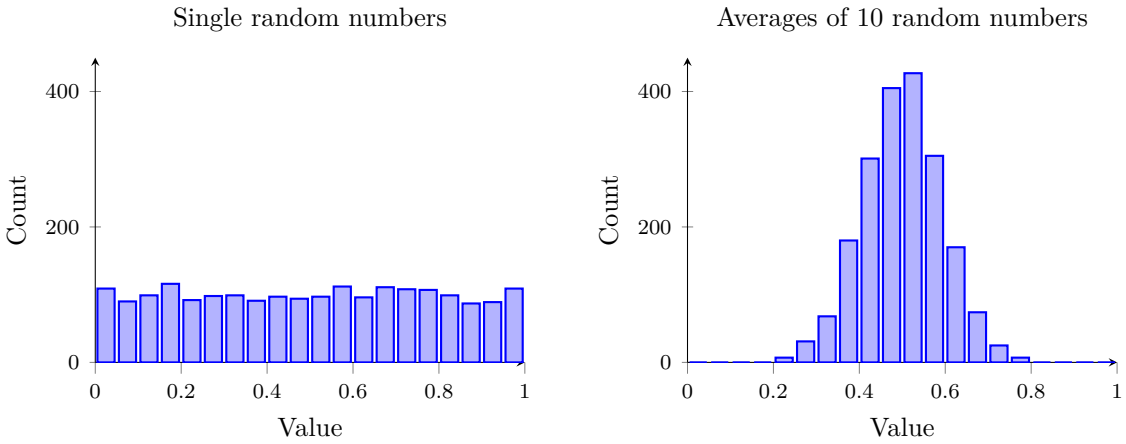
Now put the pieces together. On Day 4 we saw that the average of N samples has standard deviation $\frac{\sigma}{\sqrt{N}}$, which is the SEM. The central limit theorem says that average is roughly normal. So applying the 68-95-99.7 rule to the *average*, about 95% of the time the mean of our data falls within $2 \cdot \text{SEM}$ of the true mean. (In the simulation above, the averages have standard deviation about 0.091, which matches $\frac{\sigma}{\sqrt{10}} \approx 0.091$ for a uniform distribution. As a sanity check, they are also centered at 0.5.)

This is why we draw error bars at

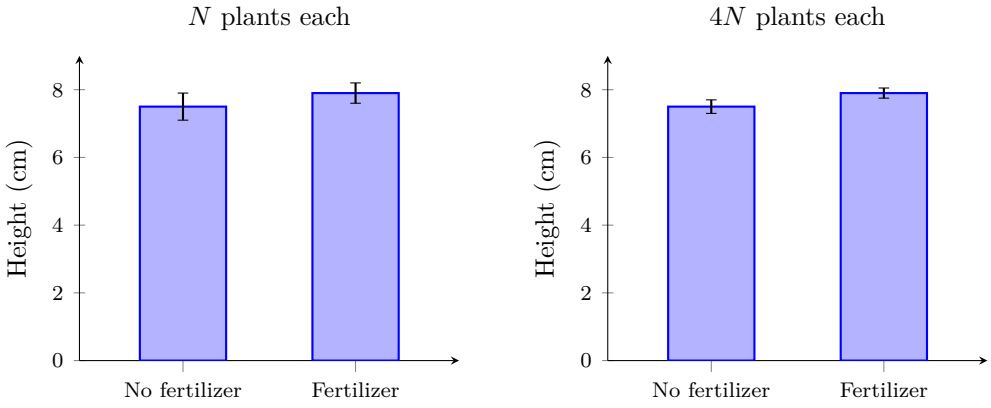
$$\bar{x} \pm 2 \cdot \text{SEM}.$$

The more precise multiple for exactly 95% is the 1.96 we used on Day 4. If the error bars of two groups do not overlap, we have evidence that the groups truly differ, and not just by luck. A bar chart with error bars gives both an answer to the question and a quick visual of it.

In the left chart the error bars overlap (7.1 to 7.9 against 7.6 to 8.2), so we cannot conclude the fertilizer does anything; the difference could be luck. In the right chart, with four times as many



The central limit theorem in action (a simulation with 2000 data points each)



Made up plant heights with error bars $\bar{x} \pm 2 \cdot \text{SEM}$. Same means, but four times the data halves the error bars.

plants, the error bars are half as long and no longer overlap (7.3 to 7.7 against 7.75 to 8.05). Same averages, but now the difference is significant. This is the whole story of why sample size matters, in one picture.

Remark 1

When you make a bar chart, the vertical axis should start at 0. Starting it at, say, 7 makes a small difference look enormous. The Excel notes have an example of how misleading this can be.

2 Histograms

A **histogram** is a bar chart for numerical data. We split the range of values into intervals called **bins** and count how many data points fall in each one. The two plots above showing the central limit theorem are histograms. Histograms do not need error bars to be useful: they show how the data is distributed, and we can still read off roughly where it is centered and how spread out it is.

A common use is with images, which is what Discussion 2 is about. A video (like the lectures) is a sequence of images in quick succession, and each pixel of an image has a brightness, a color, and a saturation. In a black and white image only brightness matters, and tallying how many pixels have each brightness gives a histogram. There is a whole field of image analysis built on this. For example, increasing the brightness of every pixel by the same amount shifts the whole histogram to the right. You can try this with your own photos in the Pixel Data Explorer.

Remark 1

Choosing the bin width matters. Bins that are too wide hide the shape of the data (everything lands in two or three bars), and bins that are too narrow make it look spiky and random. Try a few widths and pick the one that shows the shape most honestly.

3 Scatter Plots and Lines of Best Fit

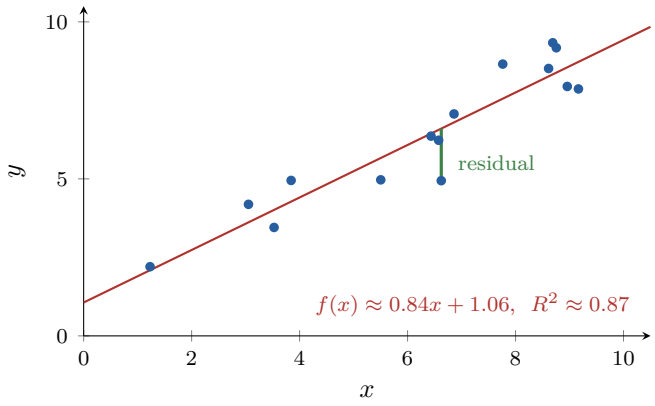
To compare two numerical variables, like age and salary, we use a **scatter plot**: one point per data point, with one variable on each axis. On Day 1 we drew a line of best fit through such a plot, and on Day 5 we measured how "line-like" it is with the correlation ρ. Here is how those connect.

Definition 3.1 (Line of best fit and residuals)

*For a line $f(x) = mx + b$ and a data point (x_k, y_k), the **residual** is $y_k - f(x_k)$, the vertical distance from the point to the line. The **line of best fit** (or **linear regression** line) is the line that makes the sum of the squared residuals, $\sum_{k=1}^{N} (y_k - f(x_k))^2$, as small as possible.*

This should feel familiar: just like variance, we square the distances so that points above and below the line do not cancel out. Excel computes this line for you as a **trendline** (see the Excel notes).

Excel also reports an R^2 value for the trendline. It is between 0 and 1, and it measures how much of the up and down in y the line accounts for. An R^2 near 1 means the points hug the line, and an R^2 near 0 means the line barely helps. For a line of best fit, R^2 is exactly the correlation squared, $R^2 = \rho_{x,y}^2$, which ties everything together. In the plot above, $\rho \approx 0.935$ and $R^2 \approx 0.87$.



A line of best fit, with one residual drawn in green

Remark 1

The slope tells you how much y changes when x increases by 1, and R^2 tells you how reliable that is. A steep line with a low R^2 is a weak result, and a shallow line with a high R^2 can be a strong one. Report both, and remember from Day 1 to be careful predicting outside the range of your data.

4 Practice

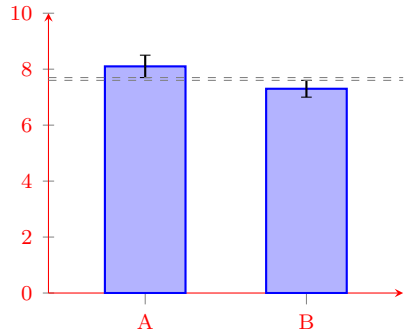
Problem 4.1 *1) Group A has mean 8.1 and SEM 0.2. Group B has mean 7.3 and SEM 0.15. Using error bars of $\pm 2 \cdot$ SEM, is the difference significant?*

2) IQ scores are designed to be normal with mean 100 and standard deviation 15. About what percent of people have an IQ between 85 and 115? Above 130?

3) A trendline has $R^2 = 0.64$ and a positive slope. What is the correlation?

Solution:

1) Group A's error bar runs from $8.1 - 0.4 = 7.7$ to 8.5, and group B's from $7.3 - 0.3 = 7.0$ to 7.6. They do not overlap, so yes. It helps to draw it:



The dashed lines at 7.7 (bottom of A) and 7.6 (top of B) show the small gap between them.

2) 85 to 115 is within one standard deviation, so about 68%. Above 130 is more than 2σ above the mean. 95% is within 2σ, leaving 5% outside, split evenly between the two tails, so about 2.5%.

3) $\rho^2 = 0.64$, so $\rho = \pm 0.8$, and since the slope is positive, $\rho = 0.8$.