The comparison of 2 groups of things is a common problem. And histograms are absolutely crucial for doing things correctly.

Suppose you want to find out which of two computers is faster. You take a given program, and run it on each computer 100 times, and record the times it takes to run the code to completion. You can then look at the histogram of “completion time” for the 2 computers:

Comparison of even 2 groups is complex, involving location, width, ...

Q: How do we handle more than 2 groups?

A: Last time we introduced the concept of the $n^{th}$ percentile (for any dist. or hist): an $x$ value with $n\%$ area to its left. Note that percentiles (or quantiles, quartiles, ...) apply to dists and hists.

Quartiles are the basis of the so-called 5-number summary of a hist (or dist), often plotted as a boxplot:

2nd quartile = $50^{th}$ percentile = median = splits data in half.

$1^{st}$ ($3^{rd}$) quartile = median of $1^{st}$ ($2^{nd}$) half.
Observations: Based on this sample, computer B is faster “on average” because its median completion time is shorter. But computer B is also more “moody” (less consistent), because it has a wider spread in completion times. Important: Note spread!!

Having said all that, one cannot conclude that computer B is faster, because these boxplots are based on a sample of size 100. We do not know what the true distribution of μ is. The true population mean (or median) of μ for each computer is somewhere in the boxplot, but we don’t know where. Given the huge overlap between the boxplots, we cannot conclude the B is faster. We cannot conclude anything! How much overlap is too much? Ans... in Ch. 7.8.

For now, just learn that every time you see a number, it’s actually a sample (of size 1), and that it’s actually a single realization of a random variable, and that the variable actually has a spread. And that’s important!
Here is an example involving many groups:

Also, as a couple of students pointed out to me, do **not** think that as a Math major you are doomed to do poorly. This info should encourage you to work harder!

I'm not showing you test 2!

Class discussion!

Note that the discussion involves comparing the whole box plots, **not** comparison of the 5 numbers one by one.

In summary, (comparative) box plots form a powerful tool of visually comparing multiple groups in terms of either data/sample from each group or their distributions.
Derivation of Binomial:

Consider \( N \) objects (population), where each object is 1 (Head, Girl, ...) or 0 (Tail, Boy, ...).

Suppose the proportion of 1's in the population is known = \( \pi \).

Now, select \( n \) (e.g., 3) of the objects (with replacement) = sample and note the value of each object.

Repeat many, many times (e.g., \( 10^8 \)).

What proportion (of the \( 10^8 \)) will be 1,1,1? 1,1,0? Etc.

Note: I’m not asking for the proportion of 1’s in each sample.

I’m asking for the proportion, out of the \( 10^8 \) trials, that are 1,1,1. Etc.

\[ X = \text{# of 1's} \]

\[ \sum_{\text{independence}} \]

\[ \begin{align*}
\text{prop of 1,1,1} & = \pi \cdot \pi \cdot \pi & = 3 \\
1,0,0 & = \pi \cdot \pi \cdot (1 - \pi) & = 2 \\
1,1,0 & = \pi \cdot \pi \cdot (1 - \pi) & = 2 \\
0,1,1 & = (1 - \pi) \cdot \pi \cdot \pi & = 2 \\
\text{Etc.} & & \\
0,0,0 & = (1 - \pi) \cdot (1 - \pi) \cdot (1 - \pi) & = 0
\end{align*} \]

\[ \begin{align*}
\text{prop} (X = 3) & = 1 \cdot \pi^3 & = \frac{3!}{3!} \\
\text{prop} (X = 2) & = 3 \cdot \pi^2 \cdot (1 - \pi) & = \frac{3!}{2!} \cdot (3-2)! \\
\text{prop} (X = 1) & = 3 \cdot (1 - \pi)^2 \cdot \pi & = \frac{3!}{1!} \cdot (3-1)! \\
\text{prop} (X = 0) & = 1 \cdot (1 - \pi)^3 & = \frac{3!}{0!} \cdot (3-0)! \\
\end{align*} \]

\[ \therefore \text{prop} (X = \pi) = \frac{3!}{\pi! (3-\pi)!} \cdot \pi^x \cdot (1 - \pi)^{3-x} \]

\( \pi = 0, 1, 2, 3 \)
\[ p(x = x) = \frac{n!}{x!(n-x)!} x^n (1-x)^{n-x} \]  
(Table II)

\( x = 0, 1, 2, \ldots, n = \# \text{ of } 1's \text{ out of } n \)

This is the mass function, \( p(x) \), of a binomial variable \( x \).

E.g. \( x = \# \text{ of heads out of } n \text{ tosses} \)

Because we derived the above expression using proportions, it follows that \( \sum_{x} p(x) = \sum_{x} \text{prop}(x) = 1 \).

Recall the connection between coin tosses and sampling:

The prob. of getting \( x \) heads out of \( n \) tosses of a coin (or \( 1 \text{ toss of } n \text{ coins} \))

\[ \uparrow \]

The prob. of getting \( x \) boys out of a sample of size \( n \).

What's \( \pi \)?

For the coin example, it's the prob. of getting a \( H \) on one toss.

In the other example, it's the prob. of drawing a boy.

i.e. The proportion of boys in the pop.

Don't confuse \( \left\{ \begin{array}{l} p(X = x) \ \ \leftarrow \ \text{Important!} \\ \text{The various \textit{proportions}}: \ \text{prop. of } 1's \text{ in each sample of } \text{size } n \end{array} \right. \)

Irrelevant! It does not show up in Binomial. It will later (Ch 7, 8).
Example 1.23 (p.56)

\[ X = \begin{cases} 3 \text{ (e.g.)} \quad 5 \text{ (e.g.)} \quad 0 \text{ (e.g.)} \quad \cdots \quad \frac{\text{# of Bads. in sample of size 100}}{100} \end{cases} \]

Assume the lots are identical, i.e., the company manufacturing the 5000 things is extremely consistent.

Then, the picture looks like this:

\[ \text{Sample 1} \quad \text{Sample 2} \quad \cdots \quad \text{Sample 10}^8 \]

What proportion of these $10^8$ lots will have $X = 0, 1, \cdots, 100$?

\[ G = \text{Good} \quad \text{Sample} = \{G, G, \cdots, G\} \]
\[ B = \text{Bad} \quad \text{Sample} = \{B, B, \cdots, B\} \]

Suppose we know the prop. of Bads, period, in the pop. = 0.5%

Then \[ P(X=x) = \binom{100}{x} \pi^x (1-\pi)^{100-x} = \frac{0.005}{7} \]

Prop. of lots with $X = 0$:
\[ (\frac{1}{100}) 0^0 (1-0.5)^{100} = \frac{1}{6058} \]
\[ = 1 \times (\frac{1}{100}) 0^1 (1-0.5)^{100-1} = 0.003 \]
\[ = 2 \times (\frac{1}{100}) 0^2 (1-0.5)^{100-2} = 0.075 \]
\[ = 3 \times \text{ Etc.} = 0.124 \]

Important Interpretation

In the long-run we expect \( \sim 60\% \) of the lots to be all good.

\( \sim 30\% \) to have 1 bad out of 100.

\( \sim 7\% \) to have 2 bads (i.e. 7\% of the lots to be 2\% defective)
We have seen the connection between binomial and the # of boys in a sample. When applied to defective gates on a chip, the binomial distribution gives:

A) proportion of all chips with n gates, that have $x$ defective gates.
B) $x < n$, that have $n$ defective gates.
C) prop. of defective gates.
D) $< n$, defective chips. $< >$.
Consider one of the two continuous variables, and one of the two discrete variables, in hw-lect 1. Make comparative boxplots for the continuous variable for each level of the discrete variable. E.g., if the discrete var. has 4 levels, then you need to show 4 boxplots for the cont. var. all on the same plot, side-by-side. Interpret.

One piece of info. That boxplots don’t convey is sample size. Let’s “run some simulations” to explore that issue. To that end:

a) take a sample of size 20 from a normal with \( \mu = 0, \sigma = 10 \).
b) 30

c) 40

d) 50

e) 100

f) Make a comparative boxplot of the 5 samples a–e.

Use the binomial mass function to show that the prob. of getting “at least 1 head out of n tosses” is \( 1 - (1-p)^n \), where \( p \) is the prob. of getting a head on a single toss.

Show work!

b) What is the numerical value of that prob. as \( n \to \infty \)?

Think about the answer you get; it’s interesting and counterintuitive.