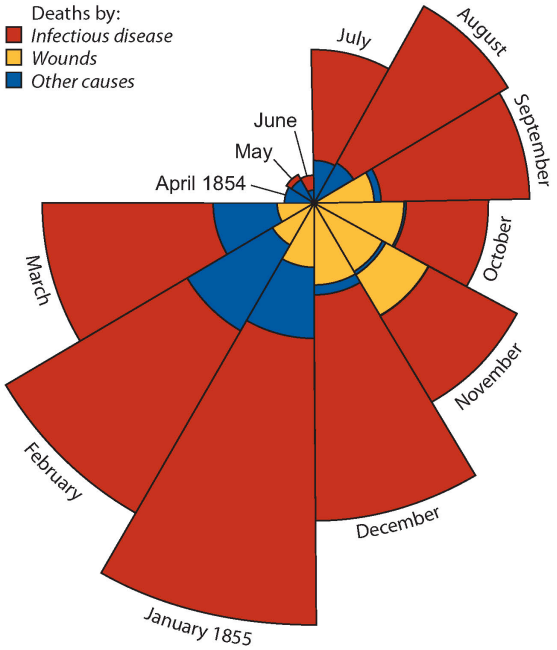
Graphics

Outline for today

- The purpose of graphs
- Types of graphs
- Examples of bad graphs
- Use tables to show patterns in data
- Principles of effective display

The purpose of graphs

- The human eye is a natural pattern detector, adept at spotting trends and exceptions. Graphs enable visual comparisons of measurements between groups and expose relationships between variables.
- Graphs are the best method available for discovering patterns in your data
- They are the principal means of communicating your results to a wider audience.



Causes of deaths in the British Army during the Crimean War (F. Nightingale 1858)
(number of deaths indicated by area of pie, measured from centre)

The purpose of graphs

Create your display so that the viewer goes “Oh!” and not “Huh?”

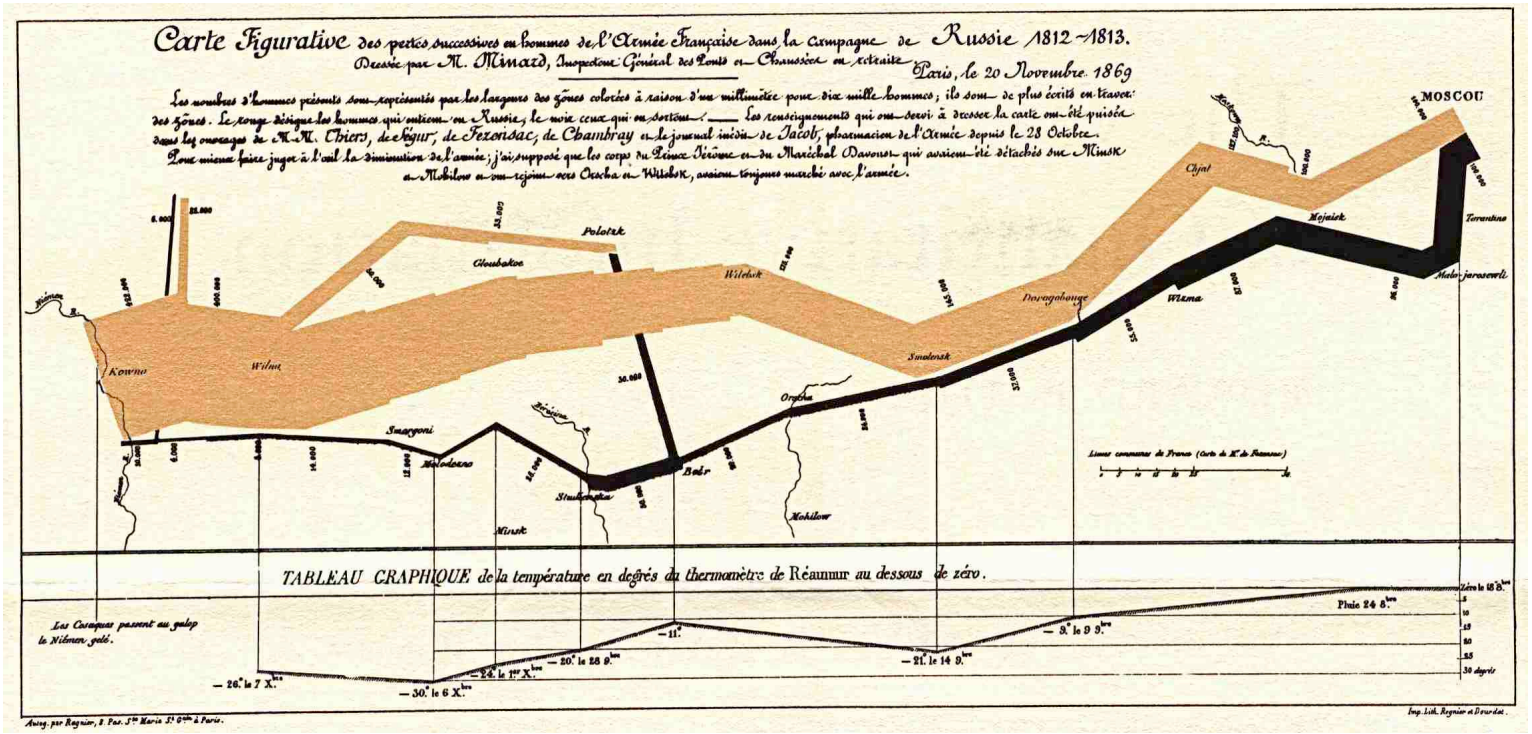
To display:

1. Frequency distributions
    - The location, spread, shape of distribution

2. Associations between variables
    - The relationship between two or more variables
    - Differences between groups in their distributions

The best statistical graphic ever drawn, according to Edward Tufte

This map by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the Polish-Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales.



Examples of types of graphs used in ecology and evolution

1. Displaying *frequency distributions*:

- Bar graphs
- Histograms

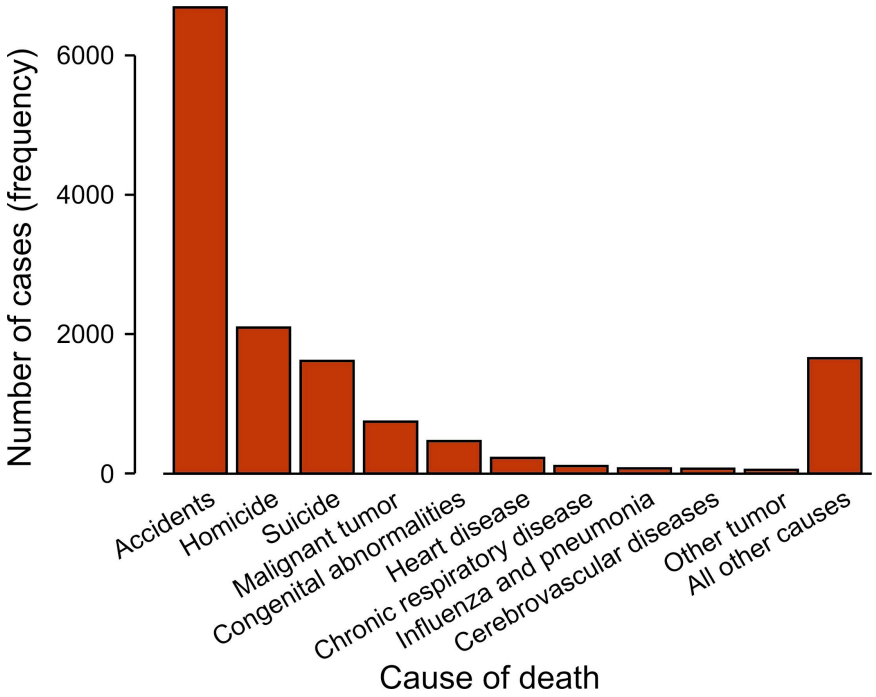
2. Displaying *associations* between variables:

- Pie chart
- Grouped bar graph
- Mosaic plot
- Box plot
- Scatter plot
- Dot plot ("stripchart" in R)

Bar graph (bar plot) - uses the height of rectangular bars to display the frequency distribution of a categorical (grouping) variable.

Features:

- zero baseline (height proportional to frequency)
- order of categories – most to least frequent is often best
- spaces between bars emphasize height

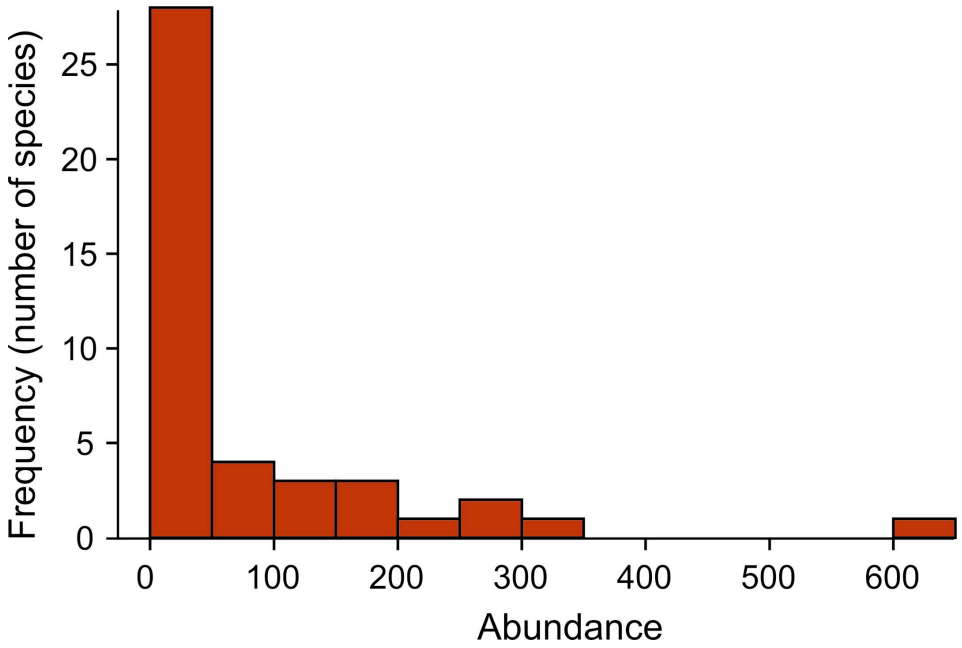


The 10 leading causes of death of Americans aged 15–19 years in 1999.

Histogram: uses the area of rectangular bars to display the frequency distribution of a numerical variable

Features:

- zero baseline (so that area proportional to frequency)
- no spaces between bars
- must choose number of bins/bin width

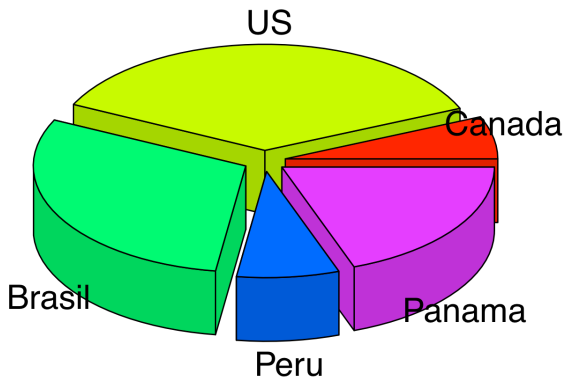


The frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. *n* = 43 species

Pie chart: the surface display association between the frequency distributions of two or more categorical variables

Features:

- frequently used.
- pretty useless. Especially, with > 3 categories and in 3D.

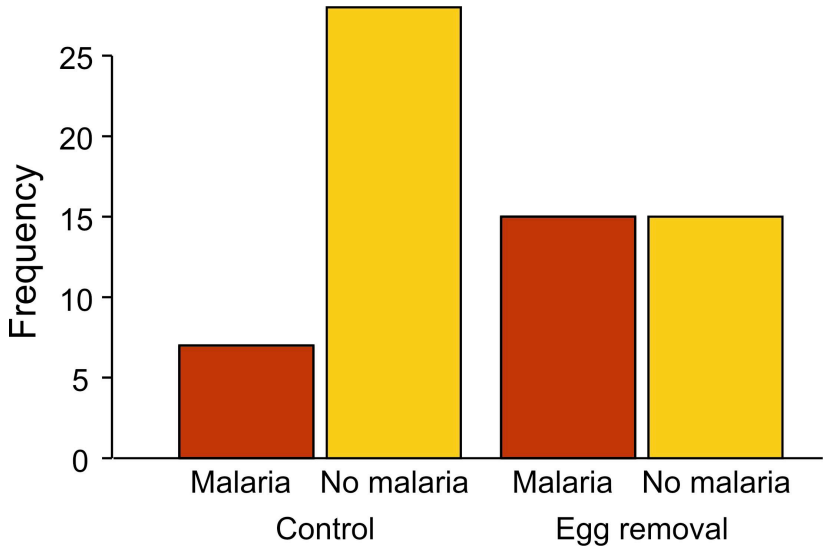


3D pie charts are so useless I even made up this data.

Grouped bar graph: uses the height of rectangular bars to display association between the frequency distributions of two or more categorical variables

Features:

- explanatory variable defines outer groups; response variable inner groups
- zero baseline (so that height is proportional to frequency)
- spacing between bars wider between outer groups

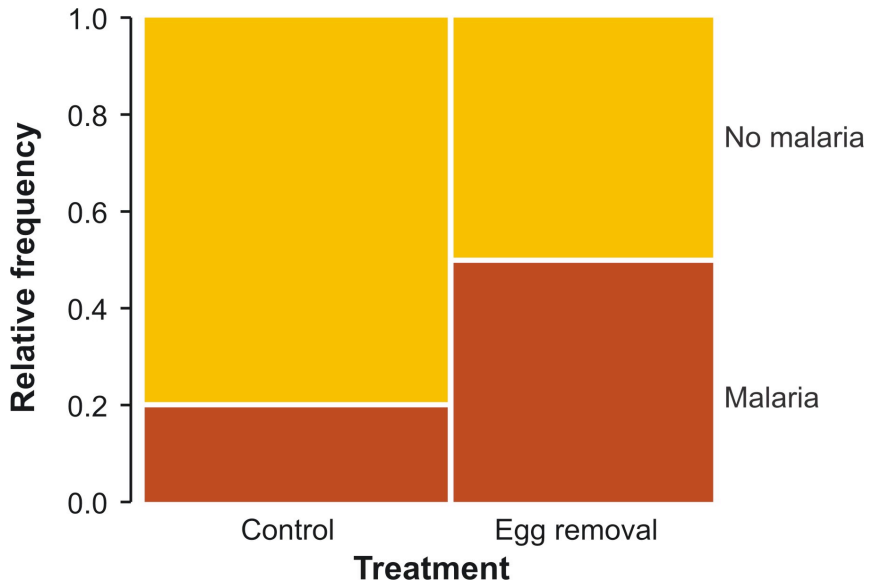


Incidence of malaria in female great tits in relation to experimental treatment. *n* = 65 birds.

Mosaic plot: uses the area of rectangles to display association between the frequency distributions of two (or more) categorical variables

Features:

- explanatory variable along horizontal axis; response variable stacked
- area proportional to frequency
- like a graphical representation of a contingency table



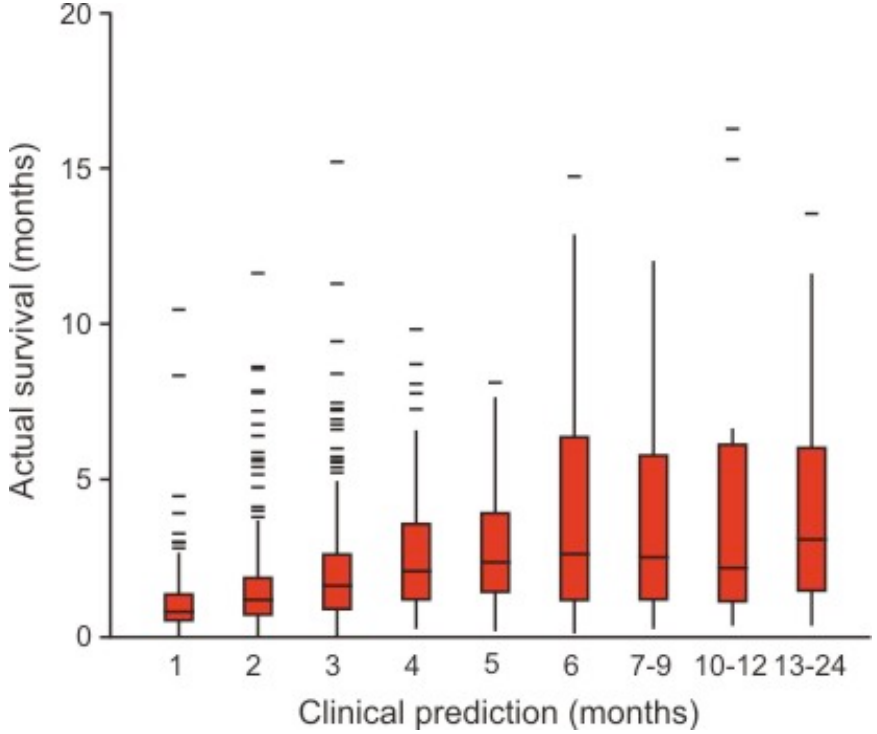
Incidence of malaria in female great tits in relation to experimental treatment. n = 65 birds.

Q: which is more successful – grouped bar graph or mosaic plot?

Box plot -- displays distribution shape for a numerical variable and its association with a categorical variable

Features:

- explanatory (grouping) variable along horizontal axis; response variable along vertical
- displays median, first and third quartile, range, and extreme observations
- non-zero baseline often ok (goal is association not magnitude or frequency)

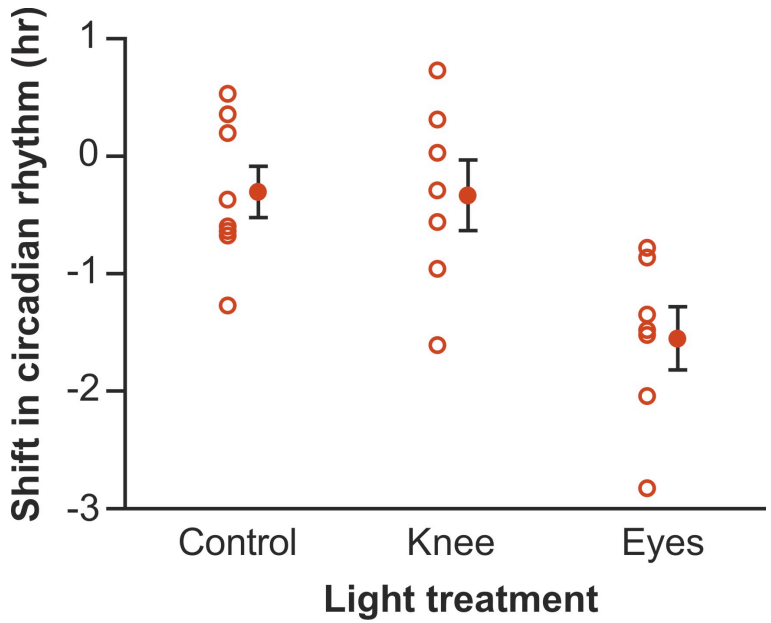


Survival times of terminally ill cancer patients with the clinical prediction of their survival times (modified from Glare et al. 2003).

Dot plot ("stripchart" in R) – displays association between a numerical variable and a categorical variable

Features:

- shows the data
- non-zero baseline often ok (goal is *association* not magnitude or frequency)
- points fill the space available

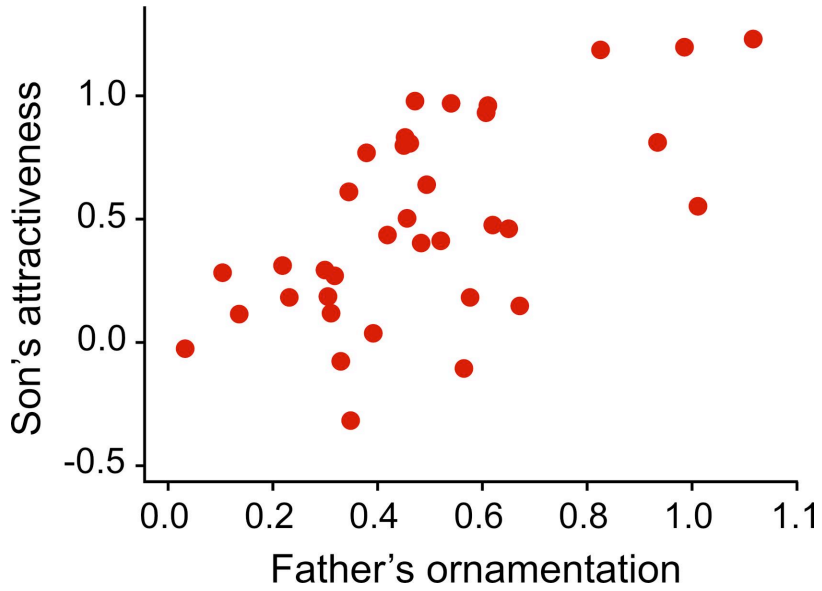


Phase shift in the circadian rhythm of melatonin production in 22 subjects given alternative light treatments (open circles). Filled dots and error bars are group means ± 1 SE. Data from Wright and Czeisler (2002)

Scatter plot -- displays association between two numerical variables

Features:

- non-zero baseline often ok (goal is *association* not magnitude or frequency)
- points fill the space available



The relationship between the ornamentation of male guppies and the average attractiveness of their sons. *n* = 36 families.

[break]

Examples of bad graphs and how to improve them

courtesy of K.W. Broman

www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

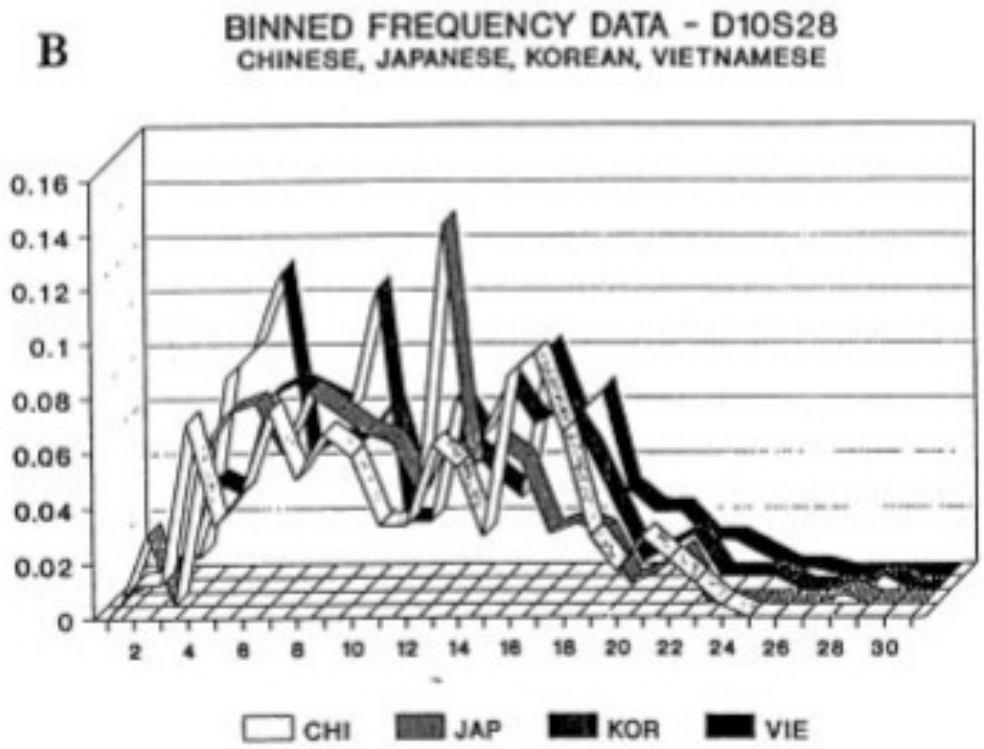


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.*

Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4

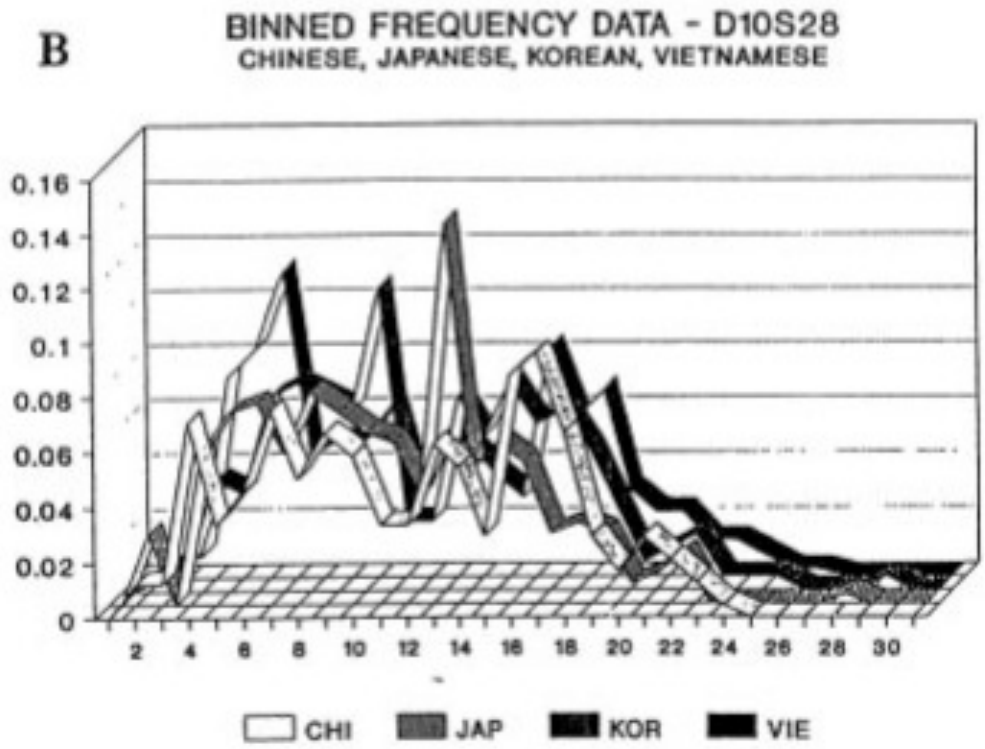


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.*

Problems:

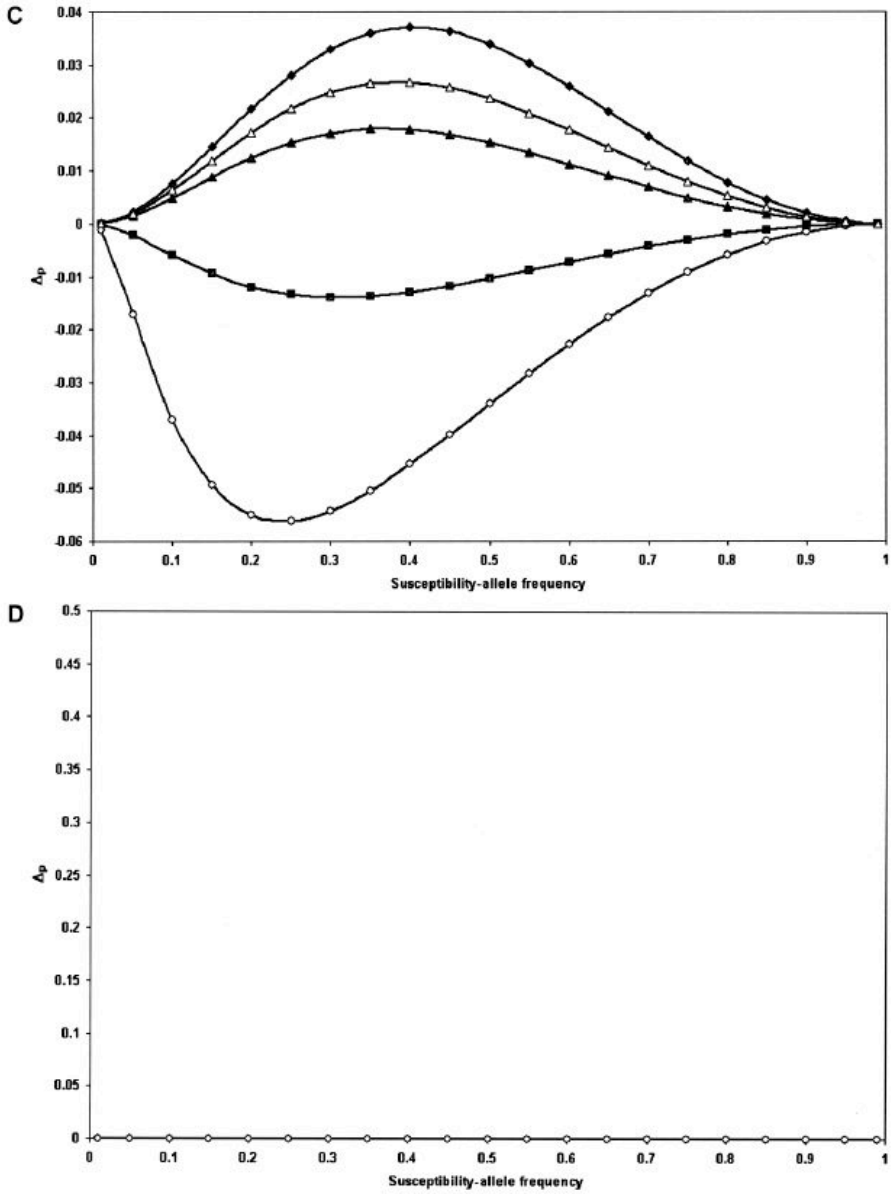
The 3-dimensional rendering of frequency distributions (“ribbons”) is confusing and unnecessary.

Solutions:

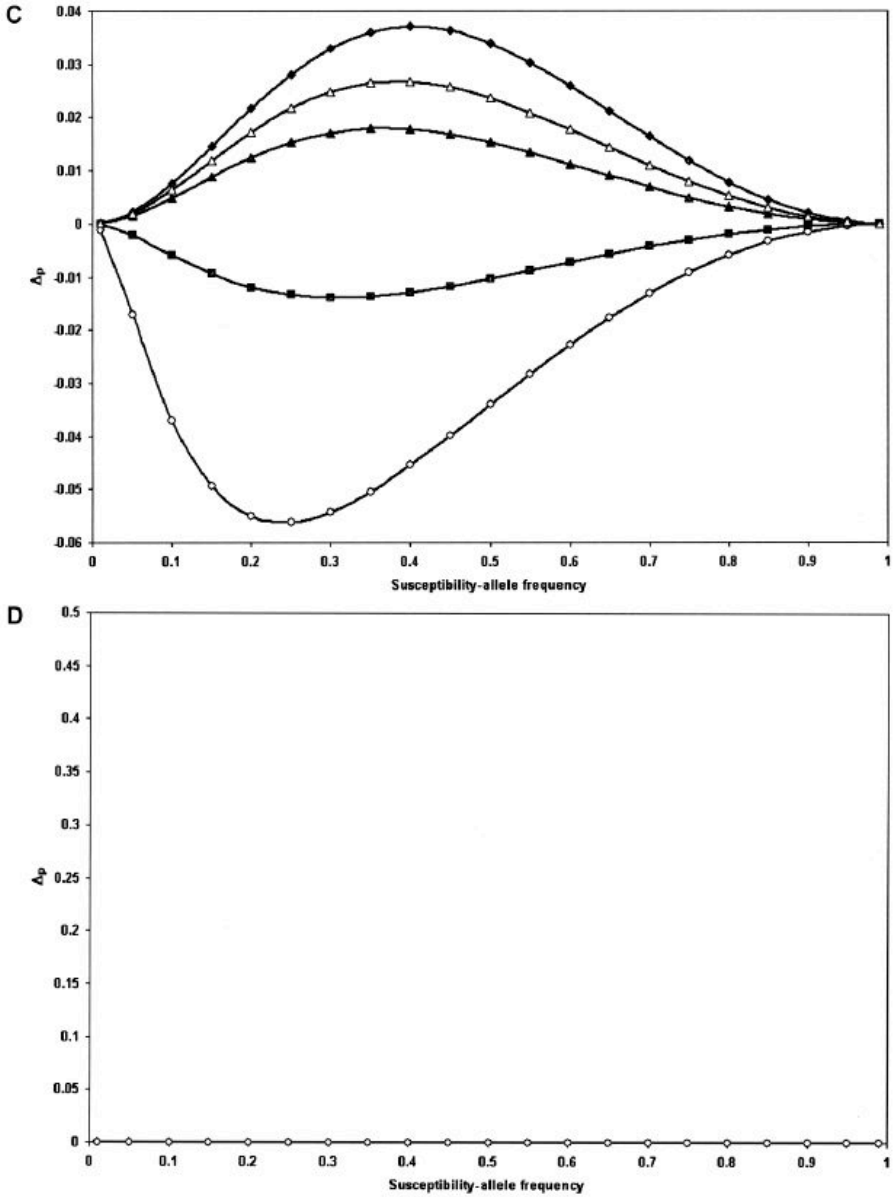
A line graph with four different line types or line colors in 2D.

Four cumulative frequency distributions in different colors or line types.

Four histograms placed in four panels arranged vertically (requires more space).



Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1



Problems:

Panel D is a waste of space; it takes a while to realize that there is any information there at all.

Solutions:

Discard D completely.

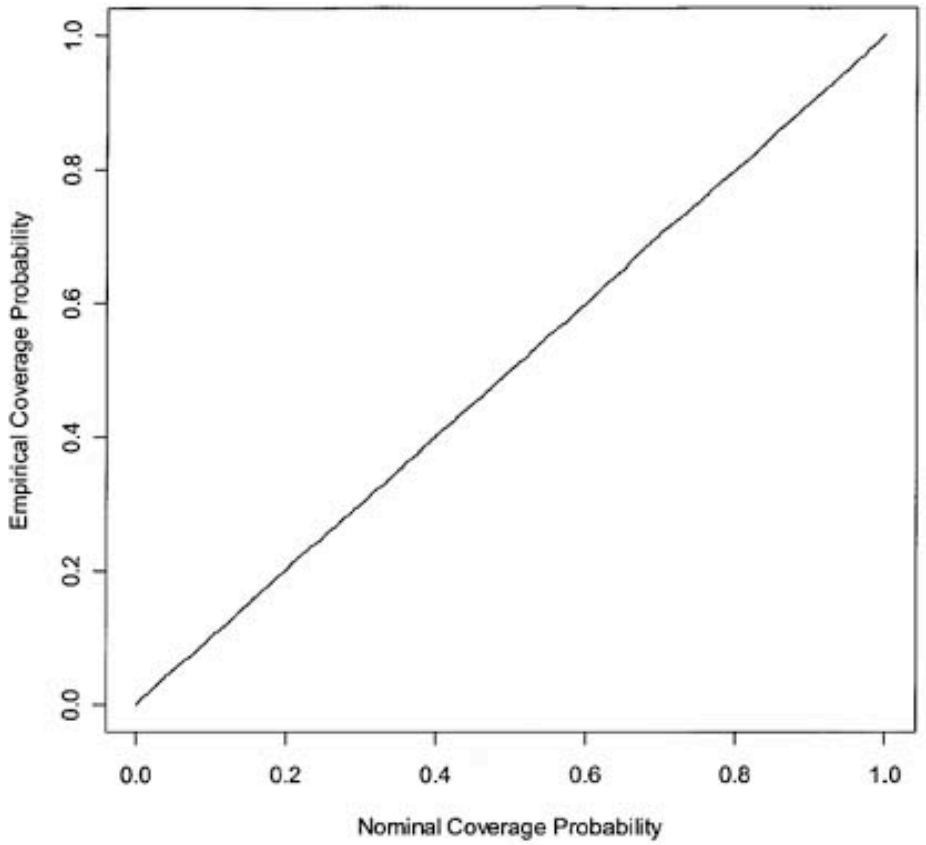


Figure 1 Empirical coverage of CIs for the relative-risk parameter β of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with $\beta = 0.35$.

Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73:1316-1329, Figure 1

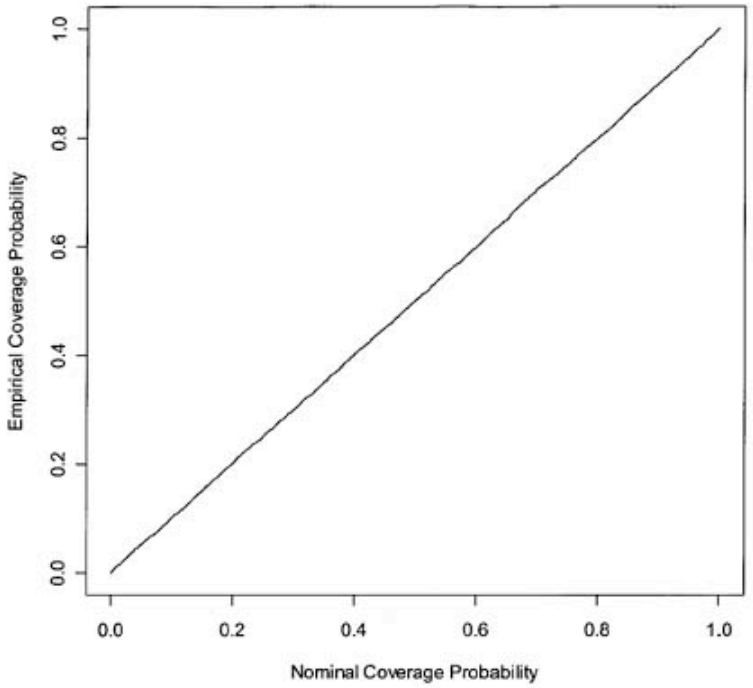


Figure 1 Empirical coverage of CIs for the relative-risk parameter β of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with $\beta = 0.35$.

Problems:

This half-page plot contains little information.

Solutions:

Plot either the percent difference between empirical and nominal (just for the region of interest)

Drop the figure and say in words that the empirical results exactly matched the nominal ones.

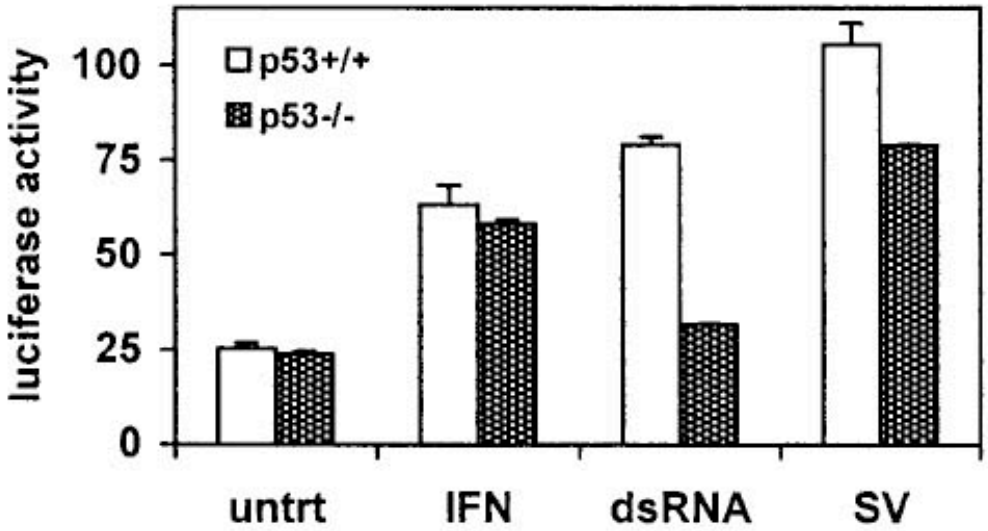


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. *J Virol* 75:7774-7777, Figure 4

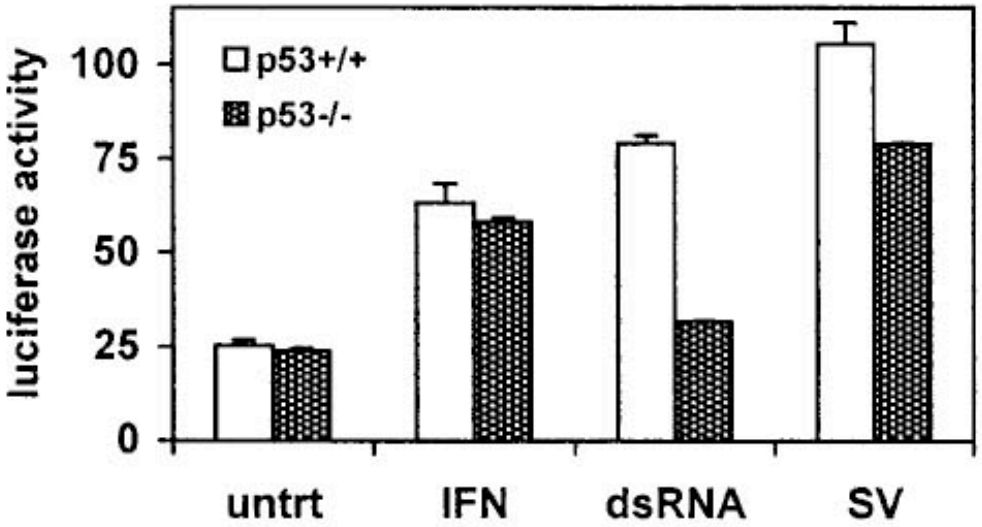


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

Problems:

The bars and little antennae represent just three data points each.

Error bars obscured by bars.

Too much ink used. All info is at the top of the bar.

Solutions:

With just three data points in each group, show the data as dots.

Replace bars with horizontal line segment to indicate means. Error bars above and below.

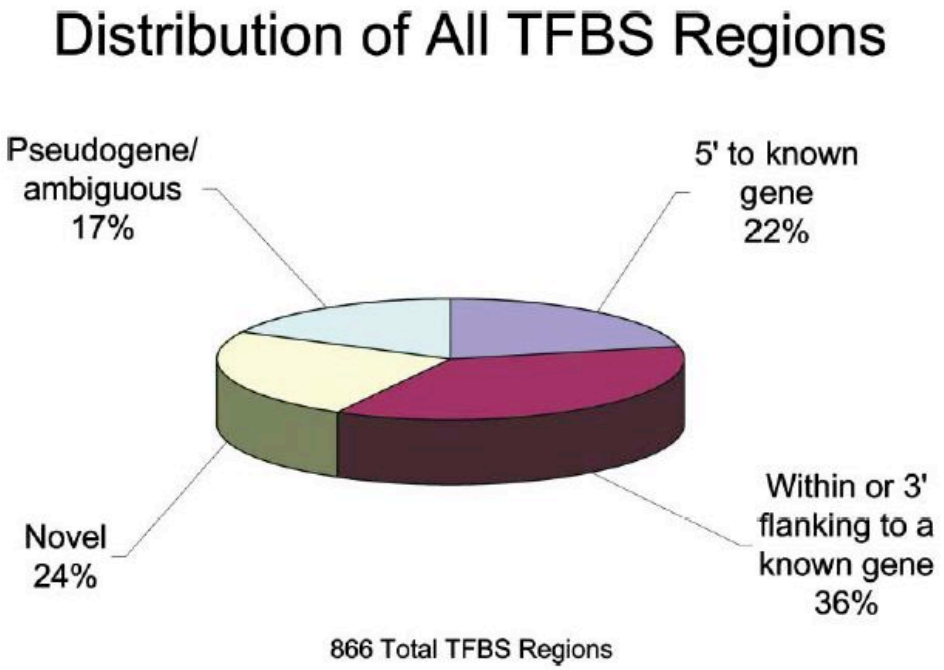


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509, Figure 1

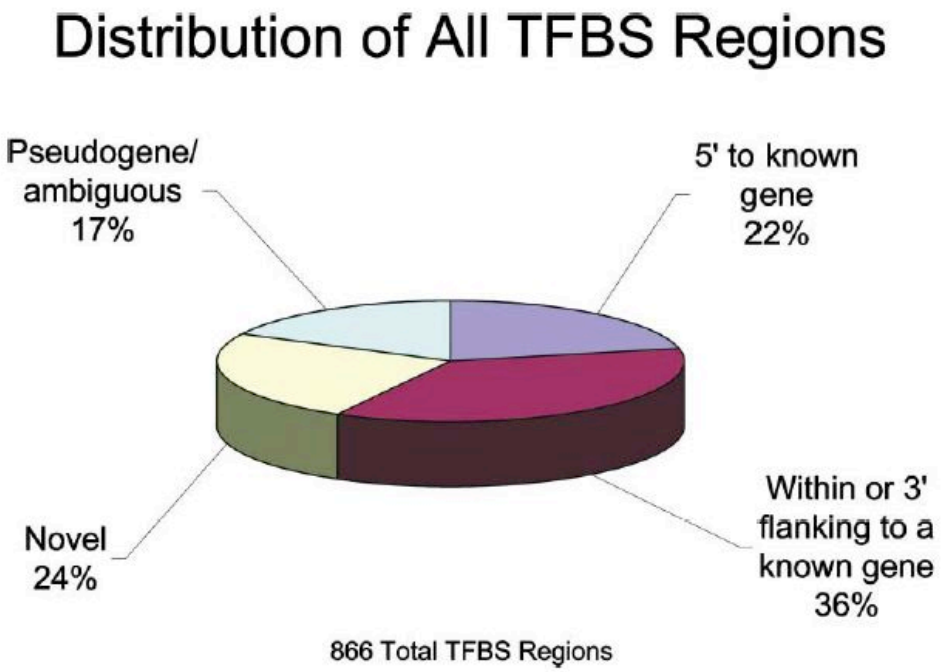


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

Problems:

3D rendering is gratuitous and confusing. Humans are poor at comparing areas in pie charts even in 2D.

Any graph that is meaningful only if the numbers are also cited must be viewed as a failure.

Solutions:

Use a 2D bar plot.

What's worse than one pie chart? A bake sale of pie charts!

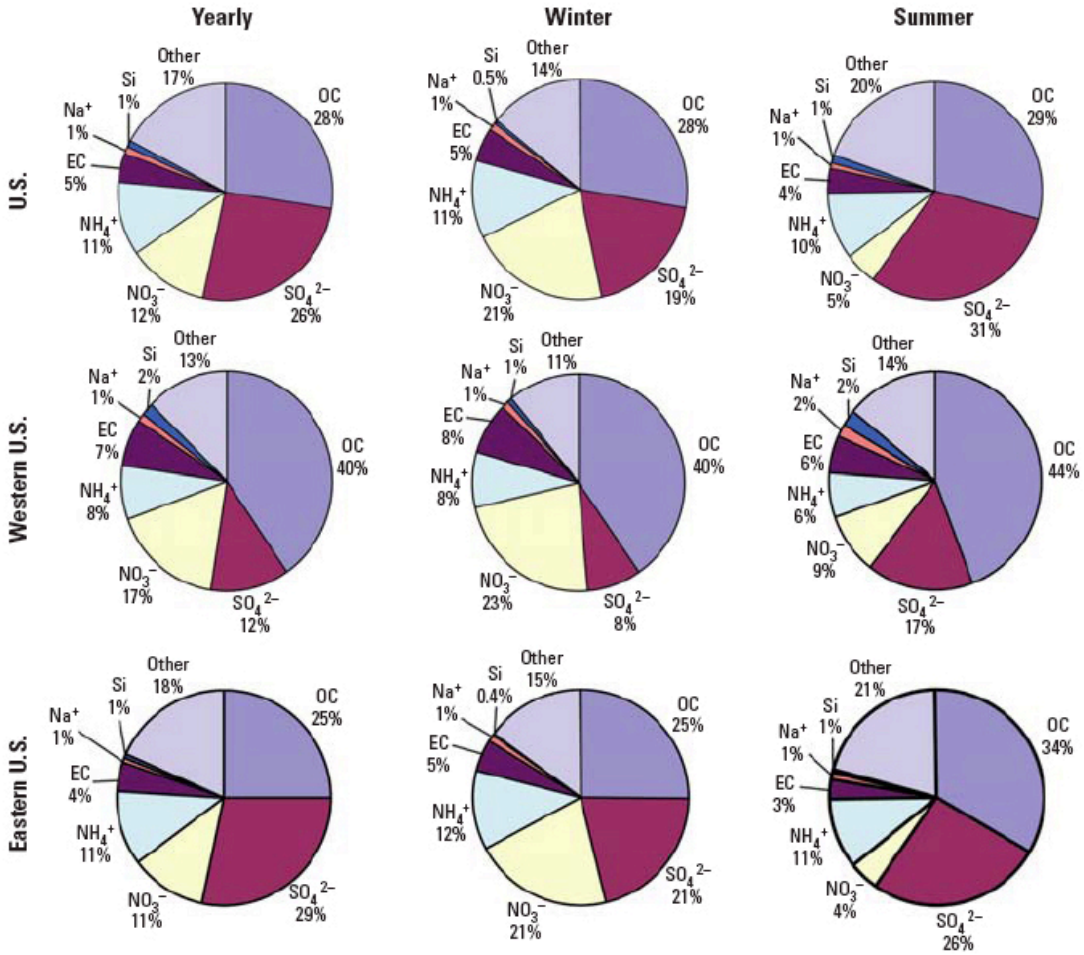


Figure 3. Percent of $PM_{2.5}$ composition by component for yearly, winter, and summer averages, by region.

Bell ML, et al. (2007) Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. *Environmental Health Perspectives* 115:989-995, Figure 3

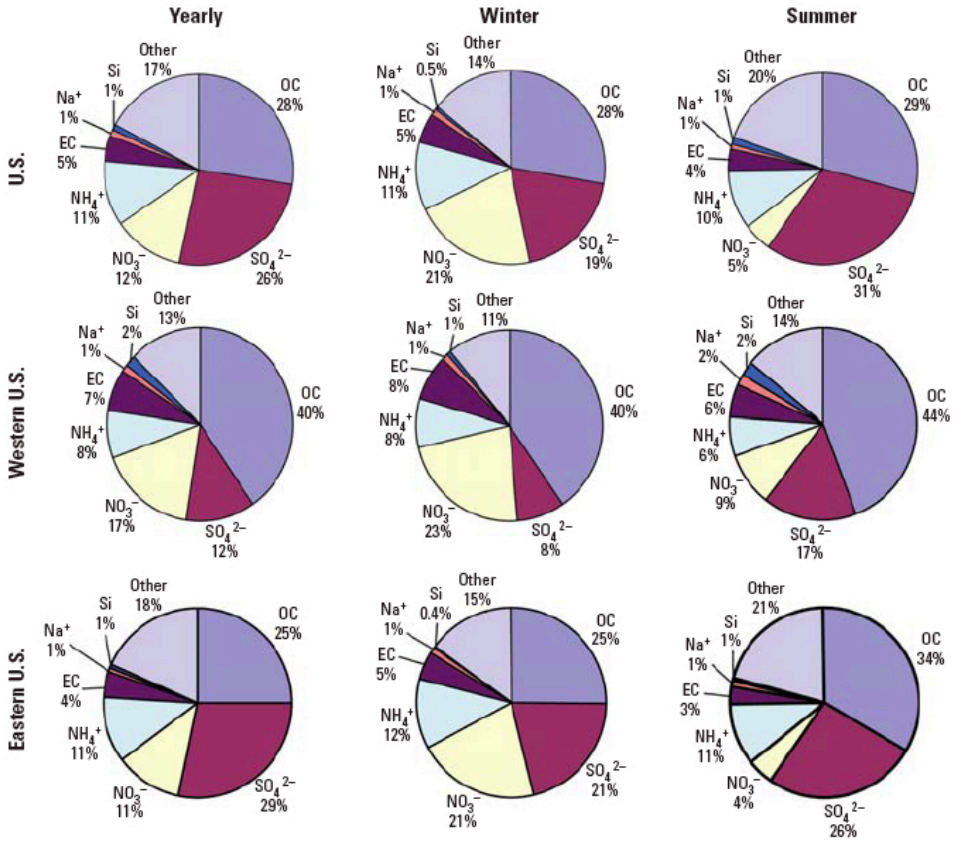


Figure 3. Percent of $PM_{2.5}$ composition by component for yearly, winter, and summer averages, by region.

Problems:

This figure is meant to show that NO_3 and SO_4 change between winter and summer and that the change is consistent across geographic region. Can you see this?

Solutions:

Use bar plots.

If all that matters is the change from summer to winter in NO_3 and SO_4, focus the graph on this aspect rather than try to display everything.

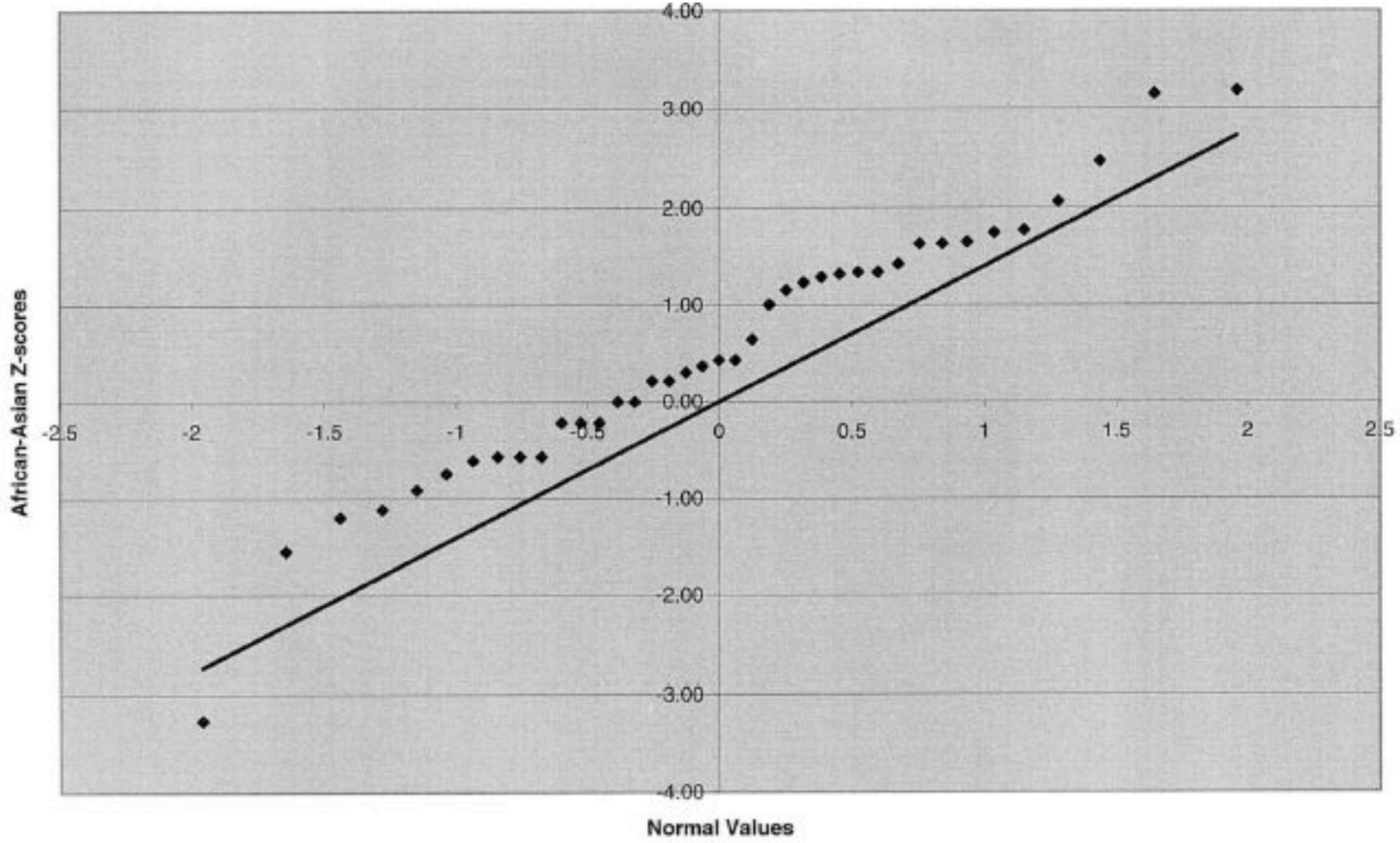


Figure 2 Q-Q plots of Z scores for telomeric interval-length differences. *a*, African Americans versus Asians. *b*, Whites versus Asians.

Jorgenson E, et al. (2005) Ethnicity and human genetic linkage maps. *American Journal of Human Genetics* 76:276-290, Figure 2

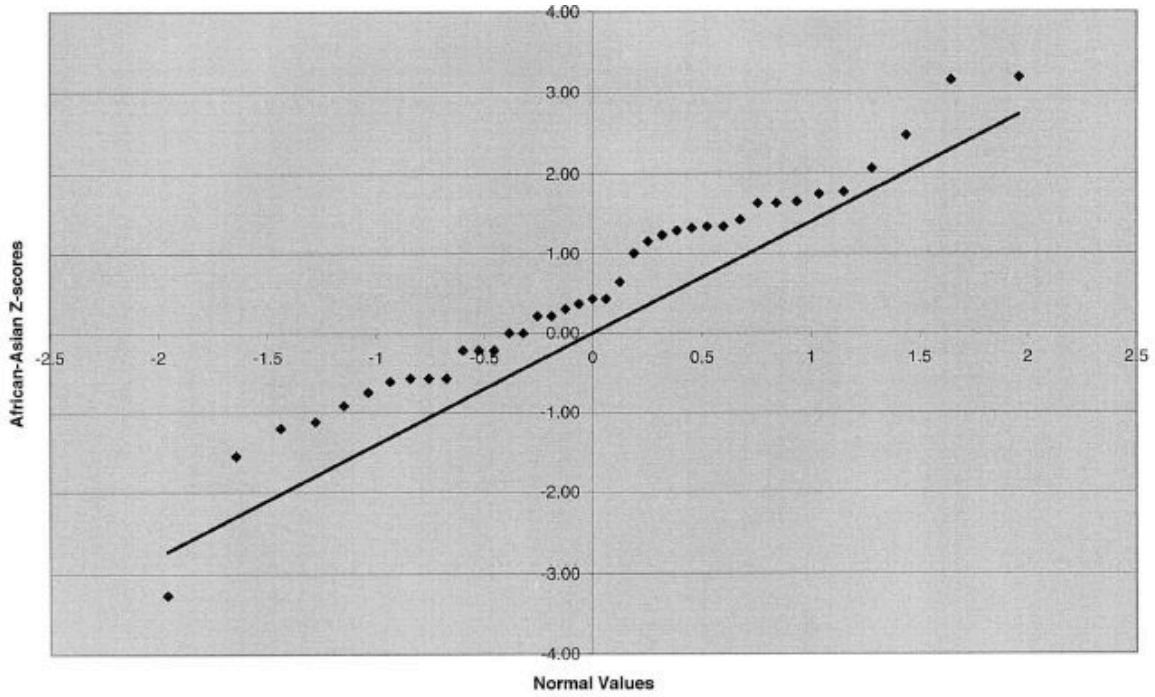


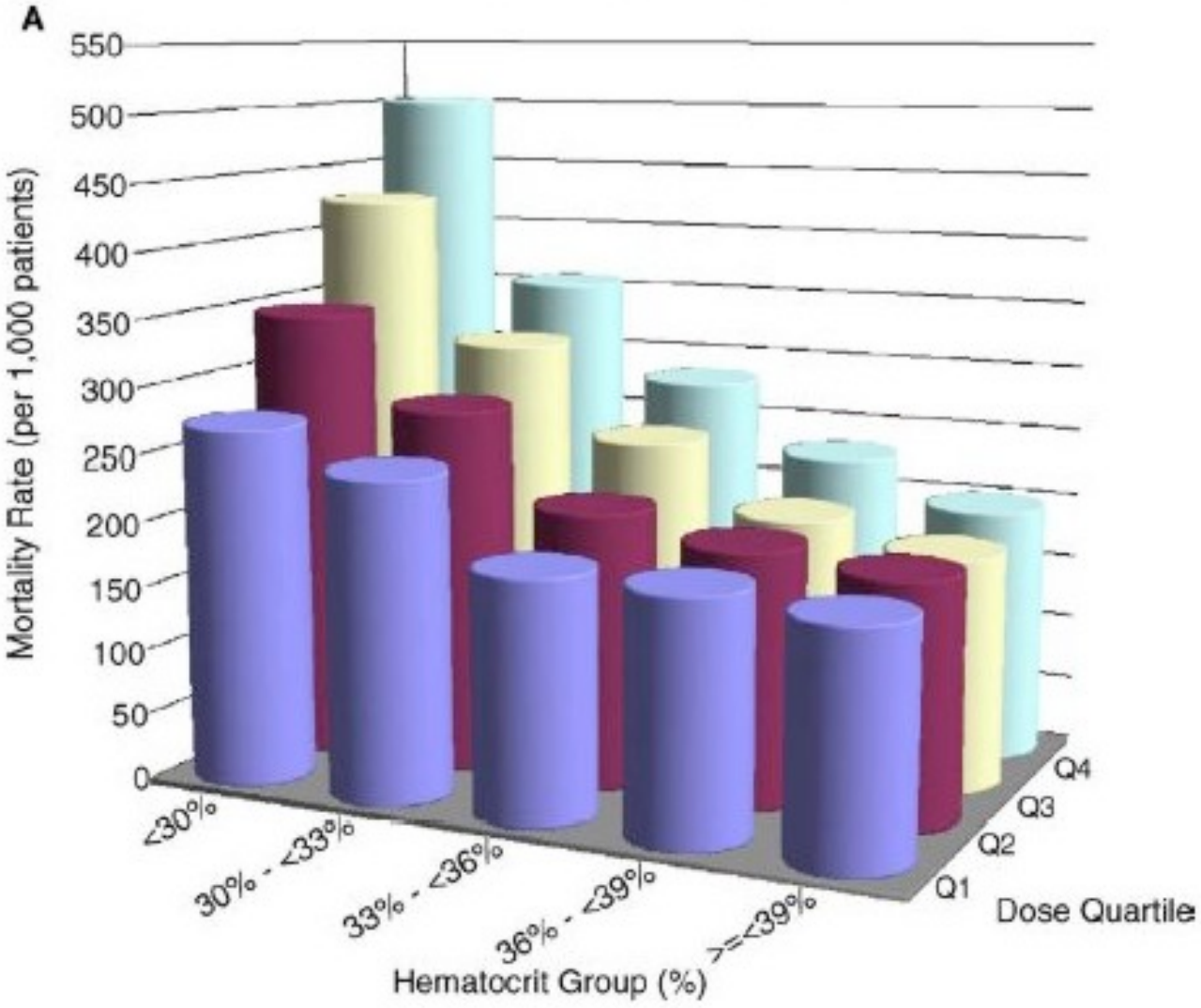
Figure 2 Q-Q plots of Z scores for telomeric interval-length differences. *a*, African Americans versus Asians. *b*, Whites versus Asians.

Problems:

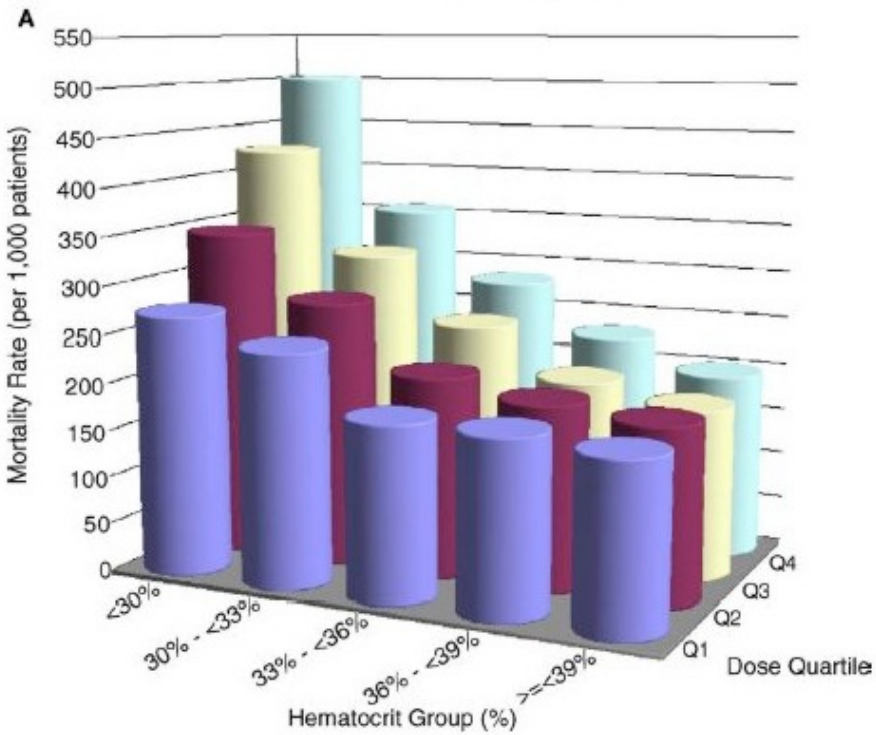
QQ plots have a lot of empty space (published figure took 2 full pages). QQ plots are for exploring not publishing.
Gray background is wasted ink, as it does not make the pattern easier to see.

Solutions:

Plot a histogram instead.



Cotter DJ, et al. (2004) Hematocrit was not validated as a surrogate endpoint for survival amoung epoetin-treated hemodialysis patients. *Journal of Clinical Epidemiology* 57:1086-1095, Figure 2



Problems:

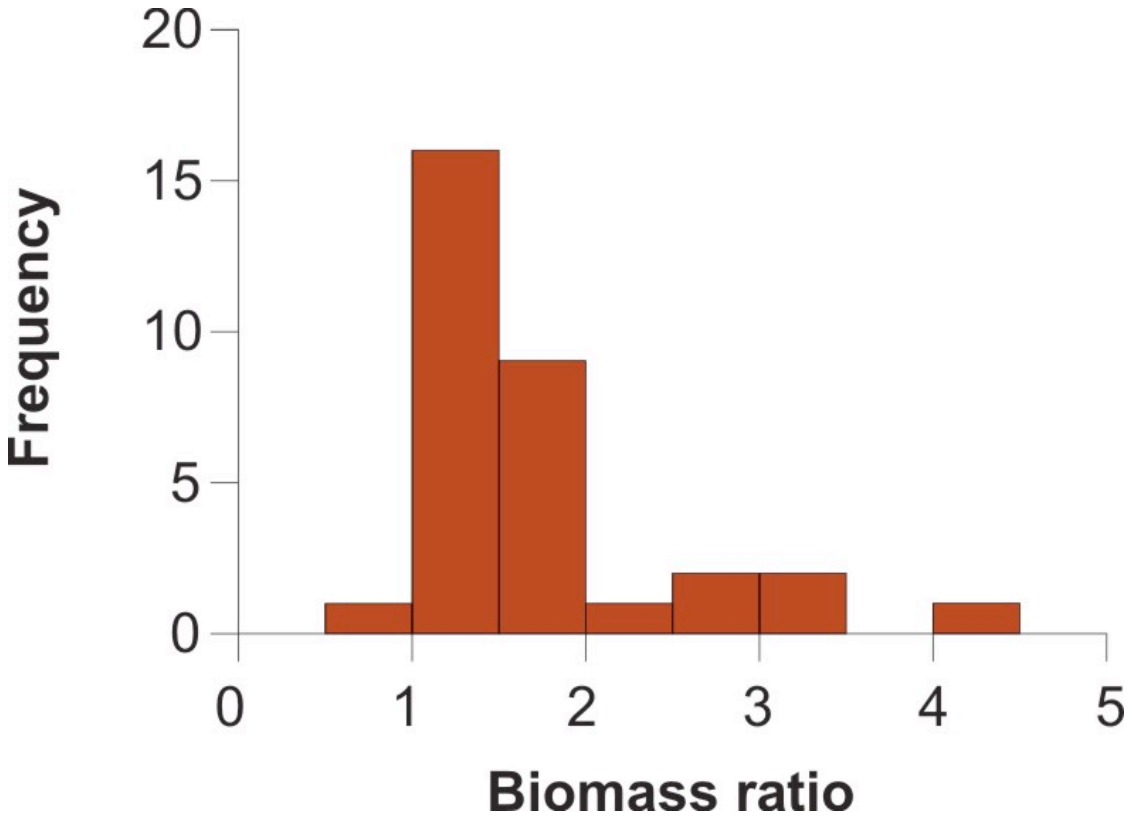
3D perspective makes it difficult to compare the areas and heights of cylinders. Embellishments = chartjunk

A lot of space (and color ink) to convey very little information.

Solutions:

Try four superposed lines of different type or color in a line plot.

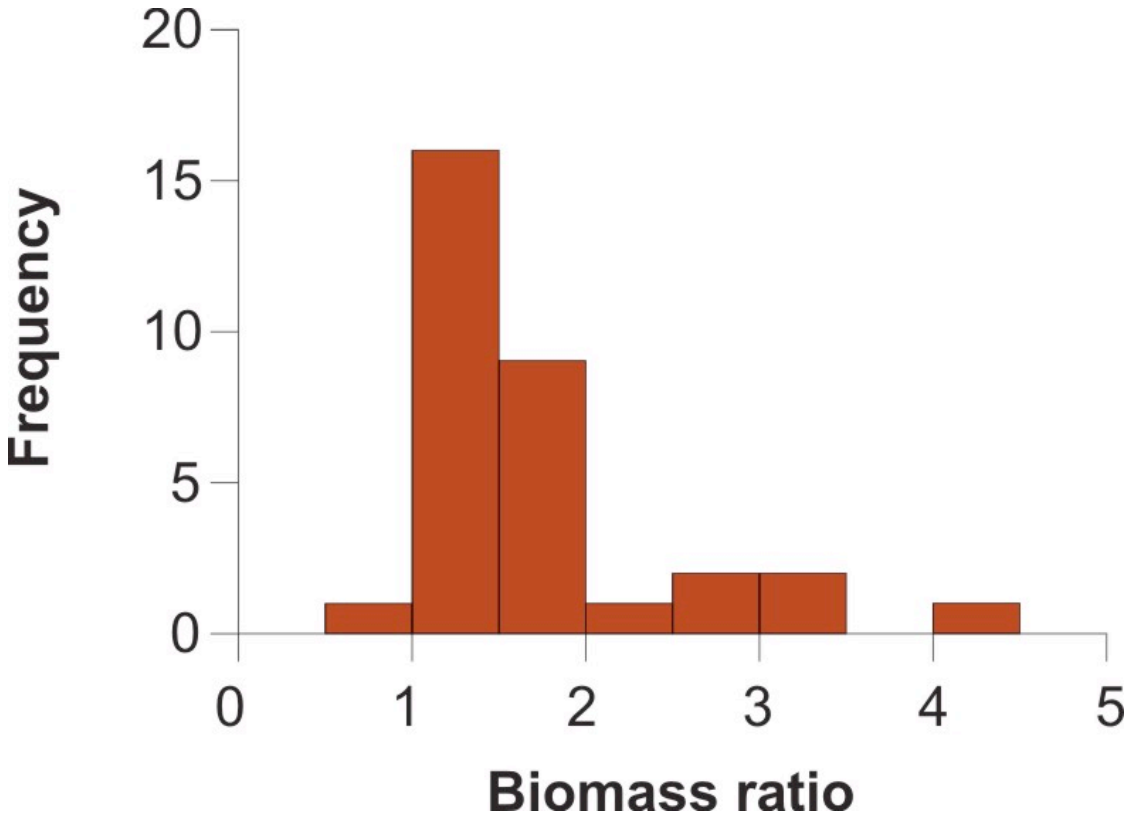
Focus effort on showing the pattern (not sure what it is in this case)



From:

Whitlock, M. C. and D. Schluter. 2008. The analysis of biological data. Roberts & Company, Greenwood Village, CO, USA, Figure 13.1-4

Biomass ratio is the total mass of all marine plants and animals per unit area of marine reserve divided by the same quantity in the unprotected control. N = 32 pairs (reserve and control). Data from Halpern (2003)



Problems:

Ratios less than 1 are sandwiched between 0 and 1, distorting magnitudes

Solutions:

Use log of ratio to give equal emphasis.

And one table

Table 5

Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

	Bias[a]		Variance[b]		95% CI[c]	
Method	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
			(M.1) $P(R=1)=0.66$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	−0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
			(M.2) logit $P(R=1)=2Y$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
			(M.3) logit $P(R=1)=2X$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	−0.04699	0.03865	0.04923	0.985	0.970
			(M.4) logit $P(R=1)=X+Y$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

[a]Bias $= (\hat{\beta}-\beta_0)/\beta_0$.

[b]Simulation variance.

[c]Confidence interval using jackknife standard error.

Paik MC (2004) Nonignorable missingness in matched case-control data analyses. *Biometrics* 60:306-314, Table 5

Table 5

Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

	Bias[a]		Variance[b]		95% CI[c]	
Method	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
			(M.1) $P(R=1)=0.66$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	−0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
			(M.2) logit $P(R=1)=2Y$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
			(M.3) logit $P(R=1)=2X$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	−0.04699	0.03865	0.04923	0.985	0.970
			(M.4) logit $P(R=1)=X+Y$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

[a]Bias $= (\hat{\beta}-\beta_0)/\beta_0$.

[b]Simulation variance.

[c]Confidence interval using jackknife standard error.

Problems:

Far too many digits. Use tables, like graphs, to display patterns not merely to store numbers.

Inconsistent number of digits.

Variance is not in same units as rest.

Unclear what main comparison of interest is. Compare methods? Compare bias with sampling error? Compare 4 missing mechanisms?

Solutions:

Fewer digits.

Use standard deviation (= standard error here) instead of variance.

Stack numbers vertically that you want most to compare rather than side-by-side or separated by other numbers.

Principles of effective display

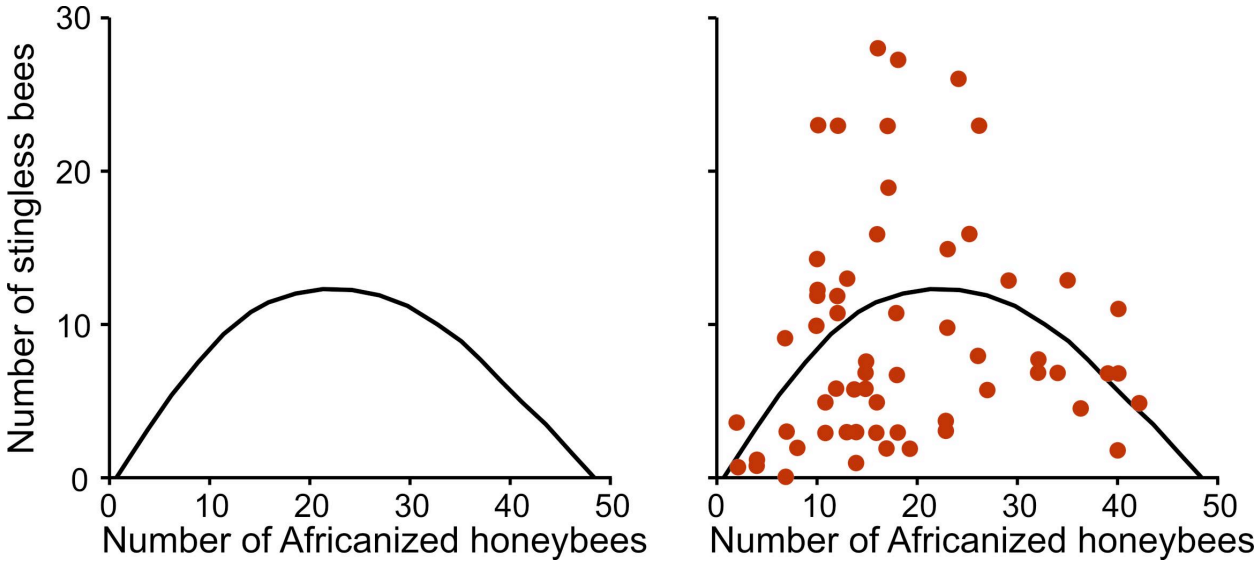
"Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" – Tufte (1983)

The following principles will help to increase the effectiveness of your graphs:

- Show the data
- Encourage the eye to compare differences
- Represent magnitudes honestly and accurately
- Draw graphical elements clearly, minimizing clutter
- Make displays easy to interpret

Show the data

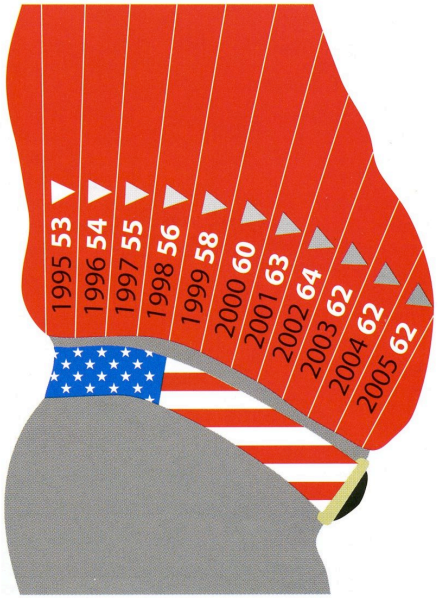
"Above all else show the data" – Tufte (1983)



The relationship between the numbers of native tropical stingless bees and Africanized honey bees on flowering shrubs in French Guiana. The data have been erased in the left panel. Redrawn from Roubik (1978).

Draw graphical elements clearly, minimizing clutter

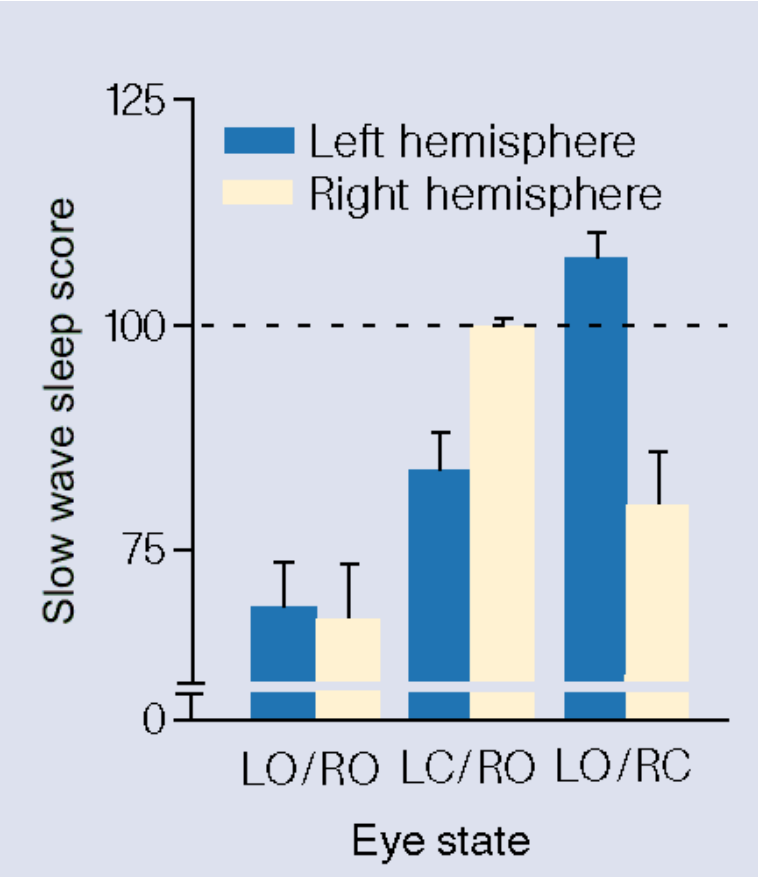
"Maximize the data-ink ratio, within reason" – Tufte (1983)



The percentage of adults over 18 with a "body mass index" greater than 25 in different years (The Economist 2006). Body mass index is a measure of weight relative to height.

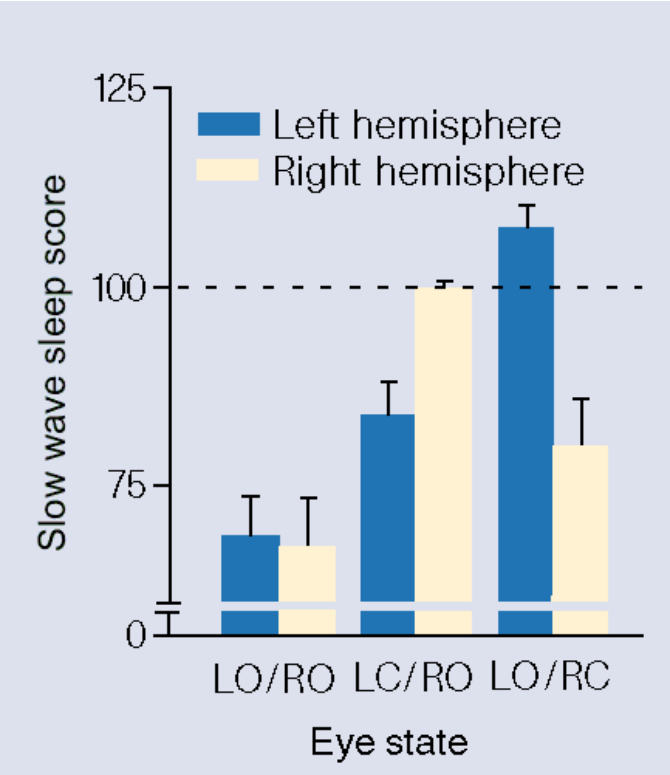
Represent magnitudes honestly and accurately

"A graphic does not distort if the visual representation of the data is consistent with the numerical representation"
– Tufte (1983)

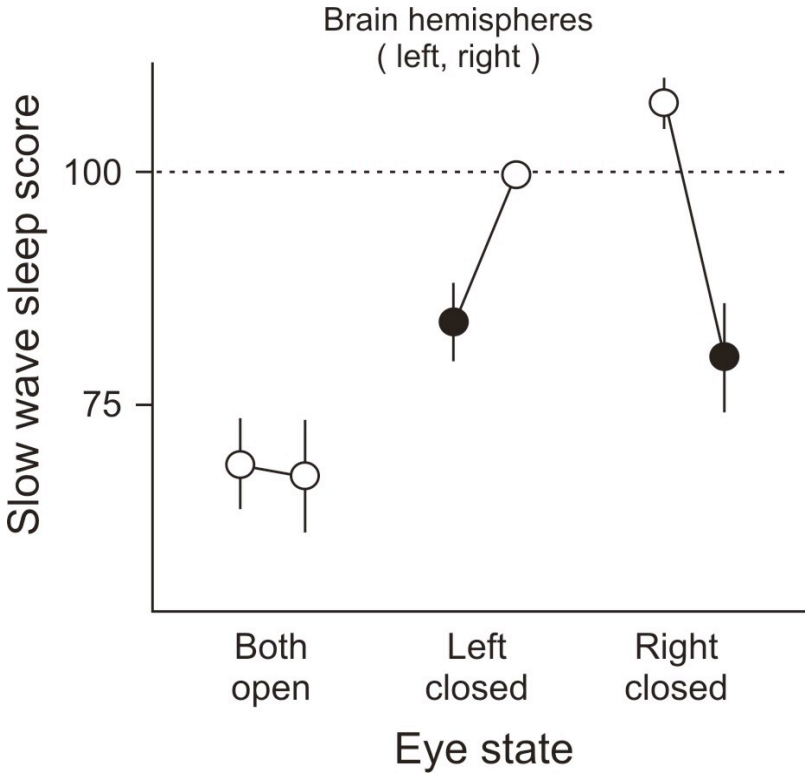


Slow wave sleep in the brain hemispheres of mallard ducks sleeping with one eye open.
From Rattenborg et al. 1999 *Nature*

Original

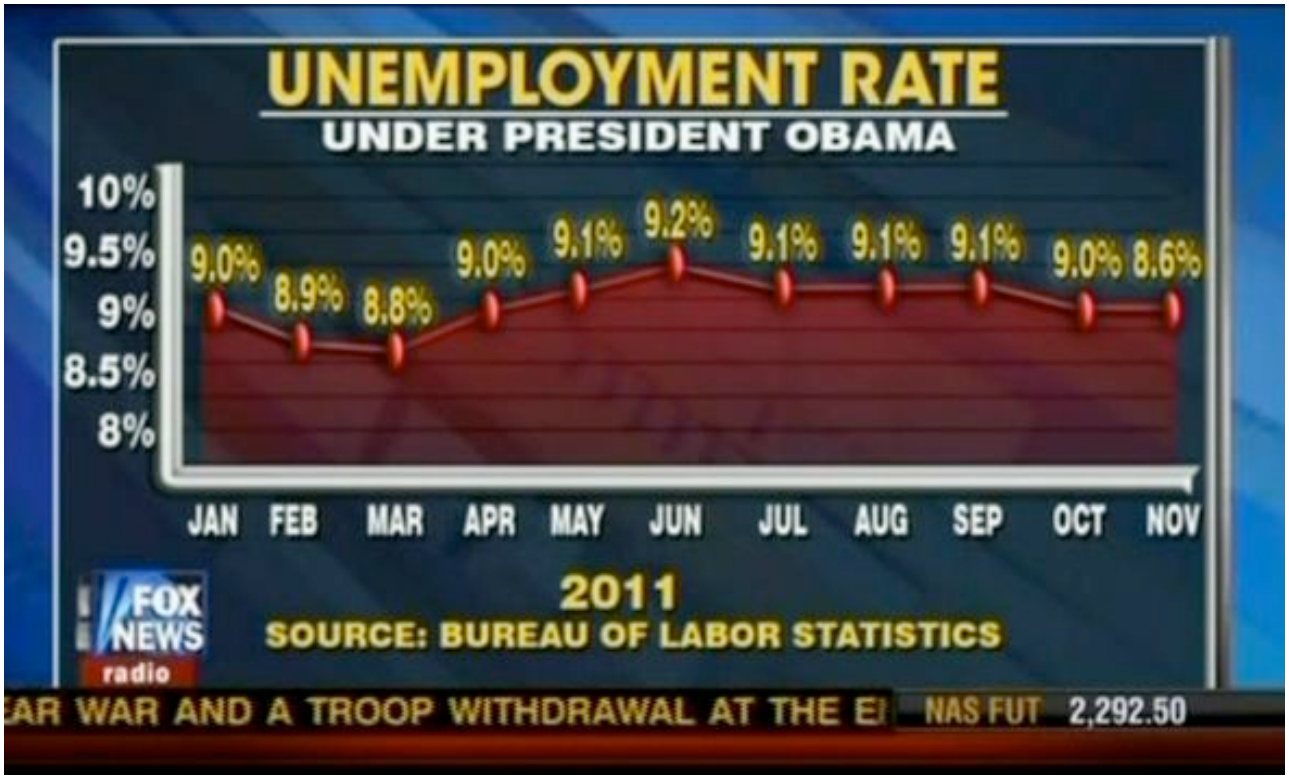


Redrawn



Slow wave sleep in the brain hemispheres of mallard ducks sleeping with one eye open.
From Rattenborg et al. 1999 *Nature*

Fox News' take on making graphs….

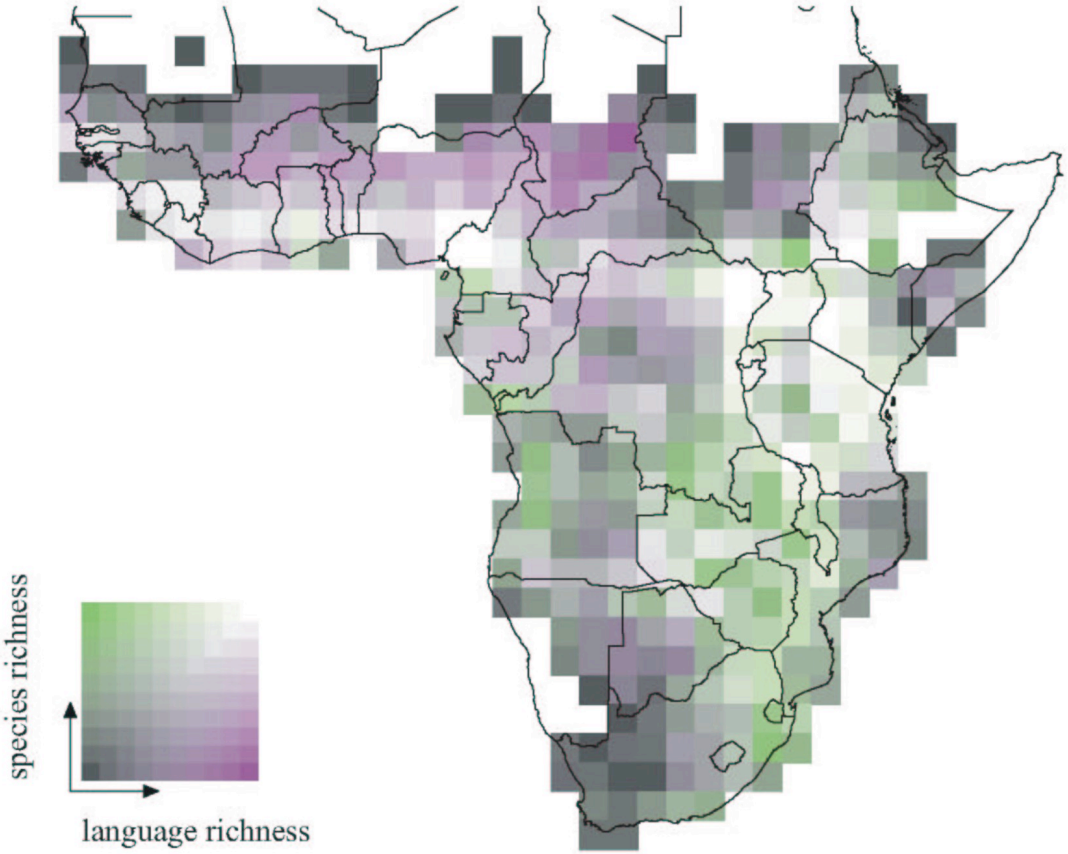


"Graphical excellence begins with telling the truth about the data" – Tufte (1983)

Let’s make Tufte proud

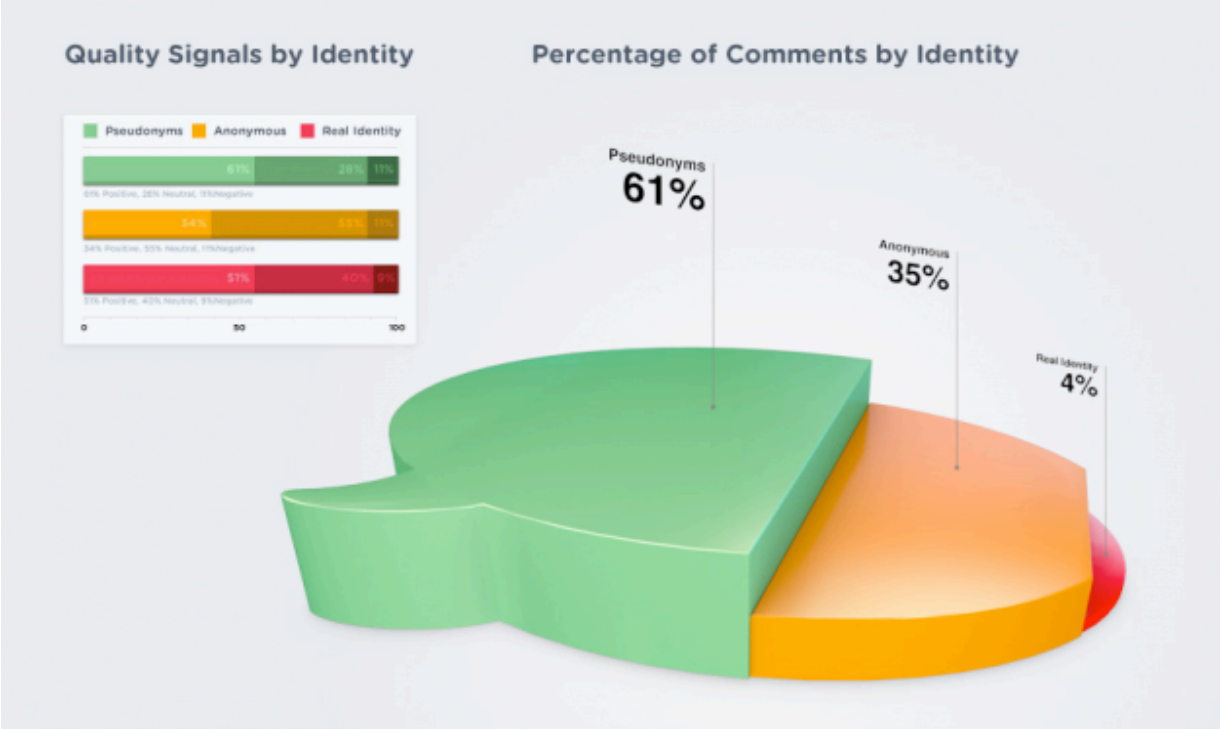
Make displays easy to interpret

"Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency"
– Tufte (1983)



Map displaying the number of bird species and the number of distinct human languages present in each square of a grid of continental Africa. Reproduced from Moore et al. (2002).

What is this?



Pie chart? Bar plot? Speech bubble?

Homework assignment 1 (due Friday, Jan 20)

(This info is repeated on "**assignments**" page on course web site)

- Find a flawed graph in a publication by your thesis supervisor (if your supervisor is flawless, pick a graph by another in your group or department).
- Explain what pattern the graph is intended to display.
- Explain how the graph is flawed in a few sentences.
- Redraw the graph in R using principles of effective display.
- Explain how your improvements display the pattern more effectively.
- Attach your R code.
- You can email it to me, preferably as a pdf file (rather than Word, etc.)
- Students from the same lab must communicate so that they don't choose the same or very similar graphs.

Discussion paper:

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187–211

Download from "**assignments**" tab on course web site.

Need:

Presenters (two) – 20-minute presentation on Thursday, January 19

Moderators (two)