# Mathematics for Policy and Planning Science

**Stephen Turnbull**

Graduate School of Systems and Information

Lecture 3: June 17, 2019

## Abstract

Introduction to modern statistics (except "big data").

# Statistics as such

- Basic use of statistics: *describe* a set of data concisely and accurately.
  - This use case is called *descriptive statistics*, as you might expect.
- A more sophisticated use case involves using statistics to determine the plausibility of scientific hypotheses.
  - This is called *inferential statistics*.
  - Until recently, inference was the primary focus of statistical theory.
- Note that it is not the calculation that determines whether a particular use is descriptive or inferential. It's the researcher's intent and conclusion that differentiates the two usages.
  - Abstract statistics like regression coefficients and Student's $t$ statistic can be used descriptively, while every statistic has a probabilistic theory when compared to a theoretical distribution, and so can be used for inference.

# Descriptive statistics

- Descriptive statistics takes the data as given.
- We can calculate *empirical distributions*, display and smooth them with *histograms*, compute *means* or *medians* to indicate the approximate "location" of the data, compute *standard deviation* (*variance*) and *range* to indicate how "spread out" or "compact" the data set is, and *skewness* to indicate direction and degree of "unbalance".
- Descriptive statistics doesn't need any probability theory directly.
  - The theory of *inferential statistics* helps to justify the use of one descriptive statistic rather than another.
  - We have already mentioned the use of the median rather than the means as one example.

# Inferential statistics

- *Inferential statistics* or *statistical inference* uses statistics to test scientific theories.
- Inference requires two kinds of models at the same time:
  - a *domain model* in mathematical or logical form
  - a *statistical model*
- The *domain model* contains the scientific, predictive content of the research. Statistics affects this model only because some mathematical models are easier to work with (*e.g.*, linear *vs.* nonlinear regression).
- The *statistical model* gives the researcher's *assumed explanation* of why the measurements don't exactly correspond to the domain model.

# Typical statistical models

- Measurement error
  - of predicted (dependent) variables: $y_t = f(x_t) + \epsilon_t$
  - of explanatory (independent) variables (also, *predictors*): $y_t = \sum_{k=1}^{m} \alpha^k (x_t^k + \epsilon_t^k)$
- Unobserved variables, usually decomposed into one or more variables:

$$\begin{aligned} \hat{y}_t &= f(x_t) + \epsilon_t \\ y_t &= \hat{y}_t \quad \text{if} \quad \hat{y}_t \geq \bar{y} \end{aligned}$$

- Random coefficients: $y_t = \sum_{k=1}^{m} (\alpha^k + \epsilon_t^k) x_t^k$
- Genuine randomness (for any of the above)
- Combinations of the above

# Statistical inference

- Consider the problem of a new vaccine of unknown effectiveness.
- We want to conduct an experiment to find out how well it works.
- There were reasons to believe that some times it would be more effective than others for reasons unrelated to the treatment (*e.g.*, in a year when few people get sick, few will catch it from them). So conduct for several years.

# Models for statistical inference

- So a model: the fraction of people from "Group $i$" who get sick is a random variable $X_i$ with support $0 \leq x \leq 1$ and continuous distribution with density $f_i(x)$. (The (cumulative) *distribution* of $X_i$ is the probability that $X_i \leq x$. You can think of the *density* of $X_i$ as the probability that $X_i = x$.)
- If we know $f_i$ for various groups $i$, then we can do comparisons (for the experiment) and predict the likelihood of an epidemic.
- We'd like to know $f_i$. Finding out is the *estimation* problem.

# Estimating $f_i$

- $f_i$ is a distribution for a continuous variable. To define $f_i$ we need a density value for every possible proportion in $0 \leq x \leq 1$—but there are an infinite number. We can smooth and interpolate with an *empirical* histogram, but it's still a lot of numbers.
- Pick an event such as $\{\omega : X_i \leq 0.2\}$ and estimate its probability.
- Take some statistic such as the expected value $\mathcal{E}[X_i]$ and try to estimate it.
- The approaches above are called *non-parametric estimation*. Alternatively, we could specify a *parametric form* for $f_i$, *i.e.*, a formula with some parameters in it, and try to estimate the parameters (*parametric estimation*). A very common parametric form is the *normal distribution* $N(\mu, \sigma^2)$. The problem is to "guess" (estimate) $\mu$ and $\sigma^2$.

# The Salk vaccine experiments

*Polio* (also called *infantile paralysis*, meaning paralysis of children) was a dreaded disease, sometimes breaking out in epidemics. Frequently affecting children, it could cause partial or complete paralysis, including preventing breathing, which resulted in death.

- The *Salk vaccine*, which injected live viruses to stimulate the immune system, was believed to be effective. But a large-scale test was needed.
- Two approaches were proposed:
  1. Select millions of children from schools. Assign 1 and 3 grade students as "test" group (gets vaccine), grade 2 as "control" (gets non-drug treatment). Ask parents of test group for permission, if not allowed, put child in control group.
  2. Select millions of students from schools. Ask parents for permission. If allowed, randomly assign to "test" group or "control" group.

# Vaccine experiment results

- With millions of students, the *law of large numbers* suggests that the results (fractions of children in each group which contract polio) should be the same.
  - Approach 1 estimated that the vaccine was effective, at about 10%.
  - Approach 2 estimated that the vaccine was effective at about 0.5%.
- More important, the *powers* of the two experiments differed. The *power* of a statistical test is the probability of a *false negative*, here, that the vaccine was observed to be ineffective even though it worked. (Note: *low* power is good!) The power of experiment 1 was about 20%, that of experiment 2 about 1%.
- Vaccination for polio has saved millions of children from contracting the disease, hundreds of thousands from paralysis, and many thousands of lives. A 20% chance of *not* using the vaccine is frightening. Why the difference?
  - The *parents* helped choose if a child got the vaccine in Approach 1, and hidden effects interfered with the experiment results.
  - In Approach 2, after given permission was given, each child was *randomly* assigned. "Hidden" effects are the same for test and control groups.

# The hidden effect

- Some parents are rich, some poor. Two direct effects:
  - Poor parents are *much less likely* to permit vaccination.
  - Rich parents provide *much cleaner* environments to their children.
- The hidden factors:
  - Very young children ($\leq$ 18 months) inherit immunity from mothers.
  - Polio virus proliferates in dirty environments.
- Indirect (hidden) effects:
  - Poor children are exposed to the virus, but don't get sick because of inherited immunity.
  - The virus stimulates their bodies to develop own, life-long immunity.
  - Poor children are quite immune to the virus, even without vaccine. Rich children are *relatively* weak against the virus, even with the vaccine.

# The mechanism

- In Approach 1,
  - there is a very high proportion of poor children in the *control* (believed high risk) group, making this group much less likely to contract polio than the general population without vaccination.
  - There is a very low proportion in the *test* (believed low risk) group, making this group much more likely to contract polio than the general population if vaccinated.
  - The difference in rates of contracting polio between the unvaccinated group and vaccinated group, while positive, is much smaller than in the general population.
- In Approach 2, these effects are absent because the randomness (with such large numbers) guarantees that the fraction of rich vs. poor children in each group is the same as the general population.
- Sampling technique is *critical* in getting accurate results.

# Types of inference problem

- *Point estimation*: a "best guess" of value of a parameter.
- *Interval estimation*: give limits for a parameter.
- *Hypothesis testing*: verify a quantitative statement.
- *Prediction*: "guessing" what $X$ will be in specific conditions.
- *Multivariate distributions*: the distribution of the policy variable (*e.g.*, number of people who get sick) depends in a statistical way on other variables ("correlation").
- *Regression analysis*: the distribution of the policy variable depends in a functional way on other variables.
- *Factor analysis*: often used in *data mining* to extract correlations with a "small number" of underlying *factors* (often interpreted as causal variables).

# Statistical Inference

- *Inference* is the practice of deducing "hidden" facts from observation.
  - People do this all the time: for example, by watching another's face, you can infer their feelings much of the time.
  - But this is "risky." For example, my uncle always looks like he disagrees with something you said—but he can't help it. Some years ago he had a stroke, and most of his face no longer moves according to his feelings.
- *Statistical inference* combines the logic of probability theory with the idea of inference.
- Here we explain *classical* (also called *frequentist*) inference, which is implemented in most statistical software. Later we discuss *Bayesian* inference, which has strong advantages, but is not as easy to use.

# Examples of Statistical Inference

- Roll a die 100 times. If all sides come up just about equally often, conclude that the die is *unbiased*.
- Roll a die 100 times. If one side comes up "too often," conclude that the die is *biased* toward that side.
- Note that these two cases may *not* be treated symmetrically!
  - The problem is that testing for bias, the *hypothesis* that the die is *unbiased* has an obvious quantitative specification:

    $$P[1] = P[2] = P[3] = P[4] = P[5] = P[6] = 1/6,$$

    and we can use it to compute the probability of any given deviation (as well as the probability of an exact match!)
  - But if the die is actually biased, the probability specification cannot be given *a priori*: any of the 6 faces might be most likely, and the deviation from equal probability need not be very great for a person who knows the bias to make a lot of money in gambling.

# The Null Hypothesis and the Alternative Hypothesis

- The *null hypothesis* is the parametrization used for calculating probabilities. It is labelled $H_0$.
- If given the null hypothesis, the probability of the observed case is high, we *accept* the null hypothesis, and *reject* the alternative hypothesis, labelled $H_1$ or sometimes $H_A$. If it is low, we *reject* the null hypothesis, and accept the alternative.
- We want to assign a decision (*accept* or *reject* $H_0$) to every observation.
  - There are some statistical tests (*e.g.*, the Durbin-Watson test for autocorrelation) where the choice is delicate, and the standard procedure actually includes undecided cases.

  However, if there are more than two possible observations, there will be many possible ways to do this.

# Warning: Change in Notation

- In the lecture, I used $\bar{p}$. I have decided to change to $\bar{n}$ in the next few slides, because $p$ is usually associated with *probability* or perhaps *proportion*, while the tests here are in terms of *counts*.
- On the other hand, $n$ is often used for counts.
- The bars in $\bar{p}$ and $\bar{n}$ mean "a fixed value."

# The Case of the Loaded (?) Die

- In the case of the die, the obvious alternative hypothesis is that *any* of the faces is too frequent. But this is not well-specified yet. The following three formula define three different sets of outcomes (*i.e.*, events), where $f(n)$ is the frequency of face $n$ in the 100 rolls, and $\bar{p}$ is "too often":

$$P[f(1) > \bar{n} \vee f(2) > \bar{n} \vee f(3) > \bar{n} \vee f(4) > \bar{n} \vee f(5) > \bar{n} \vee f(6) > \bar{n}] < \alpha \quad (1)$$

$$P\left[\sqrt{\sum_{i=1}^{n}(f(i) - \frac{100}{6})^2} > \bar{d}\right] < \alpha \quad (2)$$

$$P\left[\sum_{i=1}^{n}\left|f(i) - \frac{100}{6}\right| > \bar{d}\right] < \alpha \quad (3)$$

- You might also want to consider that if one face is *most* frequent, the opposite face is *least* frequent.
- Finally, suppose the *owner* of the die consistently bets that "3" will come up. Perhaps then you want to check if the die is loaded in her favor:

$$P[f(3) > \bar{n}] < \alpha \quad (4)$$

# Classical Inference, Significance, and Critical Values

- In the example above, we had a "distance" from "equal frequency," an event that the observed frequency was farther than that from equal frequency, and the probability of that event.
  - The event and the distance are equivalent.
  - The probability and the distance, however, actually define each other, as in this version of (4): $P[f(3) > \bar{n}] = \alpha$.
- $\alpha$ is called the *(significance) level* of the test, while the corresponding parameter (here, $\bar{n}$) is called the *critical value*. ($f(3)$ is a random variable!)
- We pick an $\alpha$ small enough that we are willing to "bet against unbiasedness", and use that to *define* the regions of rejection and acceptance of $H_0$. The *region of rejection* is sometimes called the *critical region*.
  - We "bet", because no matter how far from unbiased proportions the result is (say, "3" comes up 100 times), the probability of that happening is greater than zero if the die is actually unbiased. We must reject in this case, but *we could be wrong!*

# Error Types and Test Power

- A statistical hypothesis test generates a decision problem: Accept or reject $H_0$?
- The possible outcomes of the decision about $H_0$ can be characterized in this table:

| | Accept | Reject |
|---|---|---|
| True | No Error | Type I Error |
| False | Type II Error | No Error |

Table 1: Error types

- Thus, the *significance* $\alpha$ of a test is the probability of Type I error.
- The probability of Type II error has a name: the *power* of the test.
  - It is not obvious how to choose the power of a statistical test, because you need a specific parametric hypothesis $H_A$ to calculate the probability of rejection when $H_A$ is true, but $H_A$ in a classical hypothesis test is a range of parameter values, *i.e.*, those values bigger then the critical value.

# Regression Models

- In statistics, a *regression model* is one where there is a functional relationship between the expected value of *dependent* variable(s) and the *independent* or *explanatory* variables (also called *regressors* in the statistical context), and the actual value is "distributed around" the expected value.
- In theoretical statistics, often expressed as a *conditional expectation*.
- Here, *regress* means "to return to an original state."
  - In generic English, often deprecatory.
  - Not so in statistics, but remember that a regression is a statistical model that *assumes* a central tendency.
  - A regression model therefore attempts to "filter out" the "random" or *unexplained* variation, so the predicted outcome "returns" to the middle.

# Handle with Care

Many regression models are easy to construct and compute using software, but should be interpreted with care.

- *Type of variable* is important in interpretation. *E.g.*, when the dependent variable is *gender* with "0 = male, 1 = female", then the predicted value is often interpreted as probability. But what if $p$ is not in $0 \leq p \leq 1$?
- It's possible to transform the predicted values to the range $[0, 1]$. But there are many such transformations (in fact, the inverse of any cdf will do!) A couple are popular (logit, probit), but how do you justify them?
- Available data is often *not* the variable specified by theory, but rather a *proxy*. For example, *wage* of a worker is used by economists to represent *marginal (revenue) productivity*. Know the "quality" of your proxies.
- Regressions are subject to a large number of biases: hidden variable biases, heteroskedasticity, failure of independence.
- Regressions normally should *not* be interpreted as demonstrating causality. Domain theory *must* be used.

# The Anscombe Data Set

- The Anscombe data set is actually 4 data sets, constructed to make a point.
- Numerical calculations are very useful to provide summaries (mean, standard deviation, regression line) that are powerful aids to the researcher's intuition or presentation.
- But they can be misleading!
- With few explanatory variables, *x-y scatter plots* (including *trellis* plots) are very useful.
- With more dimensions, and a well-supported theory of a multidimensional relationship, use *residuals*.

# Summaries of the Anscombe Data Sets

| x1 | | x2 | | x3 | | x4 | |
|---|---|---|---|---|---|---|---|
| Min. | 4.0 | Min. | 4.0 | Min. | 4.0 | Min. | 8 |
| 1st Qu. | 6.5 | 1st Qu. | 6.5 | 1st Qu. | 6.5 | 1st Qu. | 8 |
| Median | 9.0 | Median | 9.0 | Median | 9.0 | Median | 8 |
| Mean | 9.0 | Mean | 9.0 | Mean | 9.0 | Mean | 9 |
| 3rd Qu. | 11.5 | 3rd Qu. | 11.5 | 3rd Qu. | 11.5 | 3rd Qu. | 8 |
| Max. | 14.0 | Max. | 14.0 | Max. | 14.0 | Max. | 19 |
| y1 | | y2 | | y3 | | y4 | |
| Min. | 4.260 | Min. | 3.100 | Min. | 5.39 | Min. | 5.250 |
| 1st Qu. | 6.315 | 1st Qu. | 6.695 | 1st Qu. | 6.25 | 1st Qu. | 6.170 |
| Median | 7.580 | Median | 8.140 | Median | 7.11 | Median | 7.040 |
| Mean | 7.501 | Mean | 7.501 | Mean | 7.50 | Mean | 7.501 |
| 3rd Qu. | 8.570 | 3rd Qu. | 8.950 | 3rd Qu. | 7.98 | 3rd Qu. | 8.190 |
| Max. | 10.840 | Max. | 9.260 | Max. | 12.74 | Max. | 12.500 |

# ANOVA for the Anscombe Data Sets

| Response = y1 | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| x1 | 1 | 27.510 | 27.5100 | 17.99 | 0.00217 |
| Residuals | 9 | 13.763 | 1.5292 | | |

| Response = y2 | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| x2 | 1 | 27.500 | 27.5000 | 17.966 | 0.002179 |
| Residuals | 9 | 13.776 | 1.5307 | | |

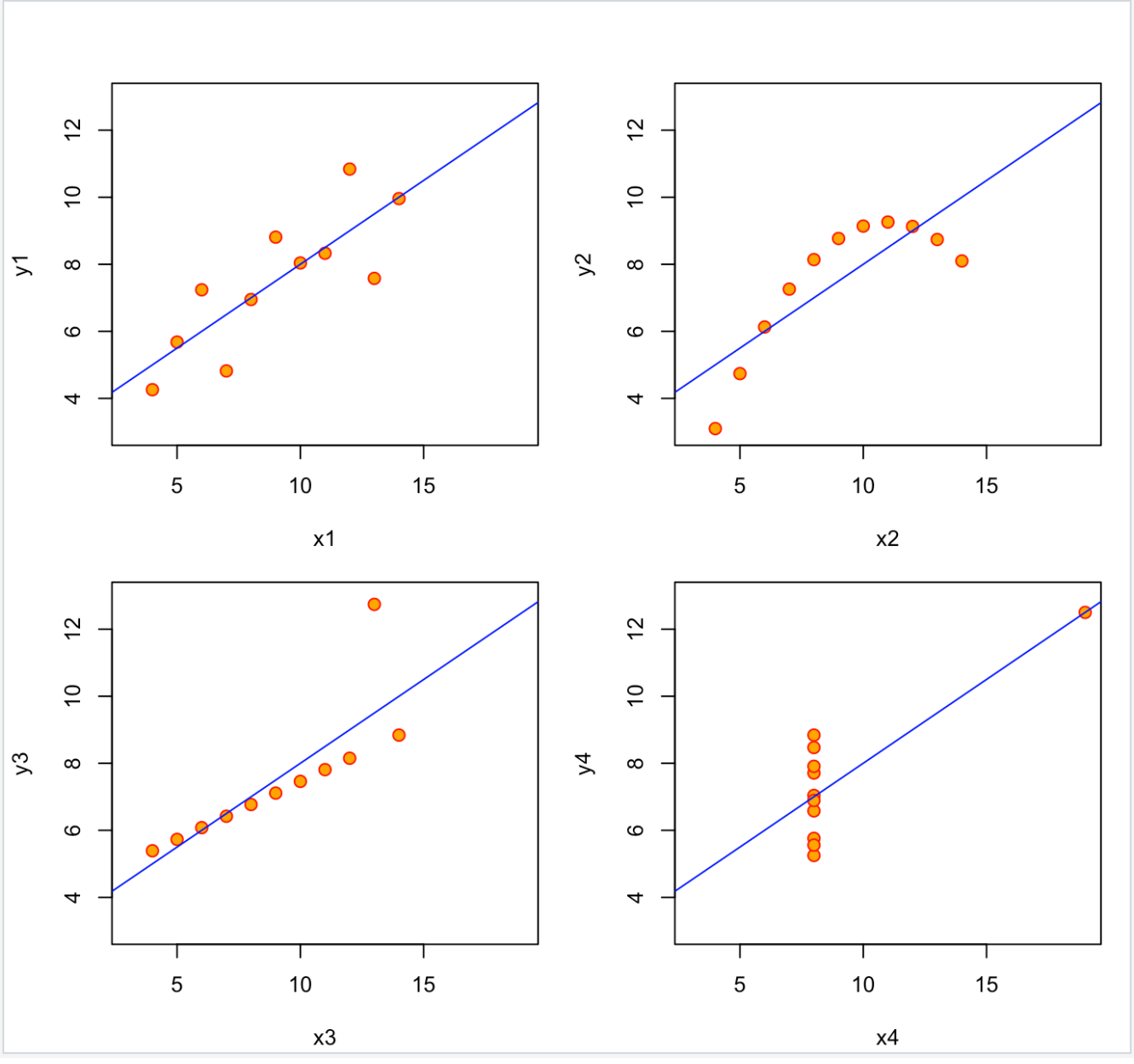
| Response = y3 | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| x3 | 1 | 27.470 | 27.4700 | 17.972 | 0.002176 |
| Residuals | 9 | 13.756 | 1.5285 | | |

| Response = y4 | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| x4 | 1 | 27.490 | 27.4900 | 18.003 | 0.002165 |
| Residuals | 9 | 13.742 | 1.5269 | | |

# Understanding the Anscombe Data Set, Part I

- The Anscombe data set shows that the same statistical summary can be produced by data which humans are likely to interpret in very different ways.
  - "Interpret" may mean "describe," or
  - re-analyze with a different formal model, or even
  - change the data.
- Visualization methods such as plots of the data can be very helpful in spotting surprising or "model-violating" patterns in a data set. See the 4 data sets plotted on the next slide. (The blue lines are the least-squares regression lines.)

# Plots of the Anscombe Data Sets



# Regressions for the Anscombe Data Sets

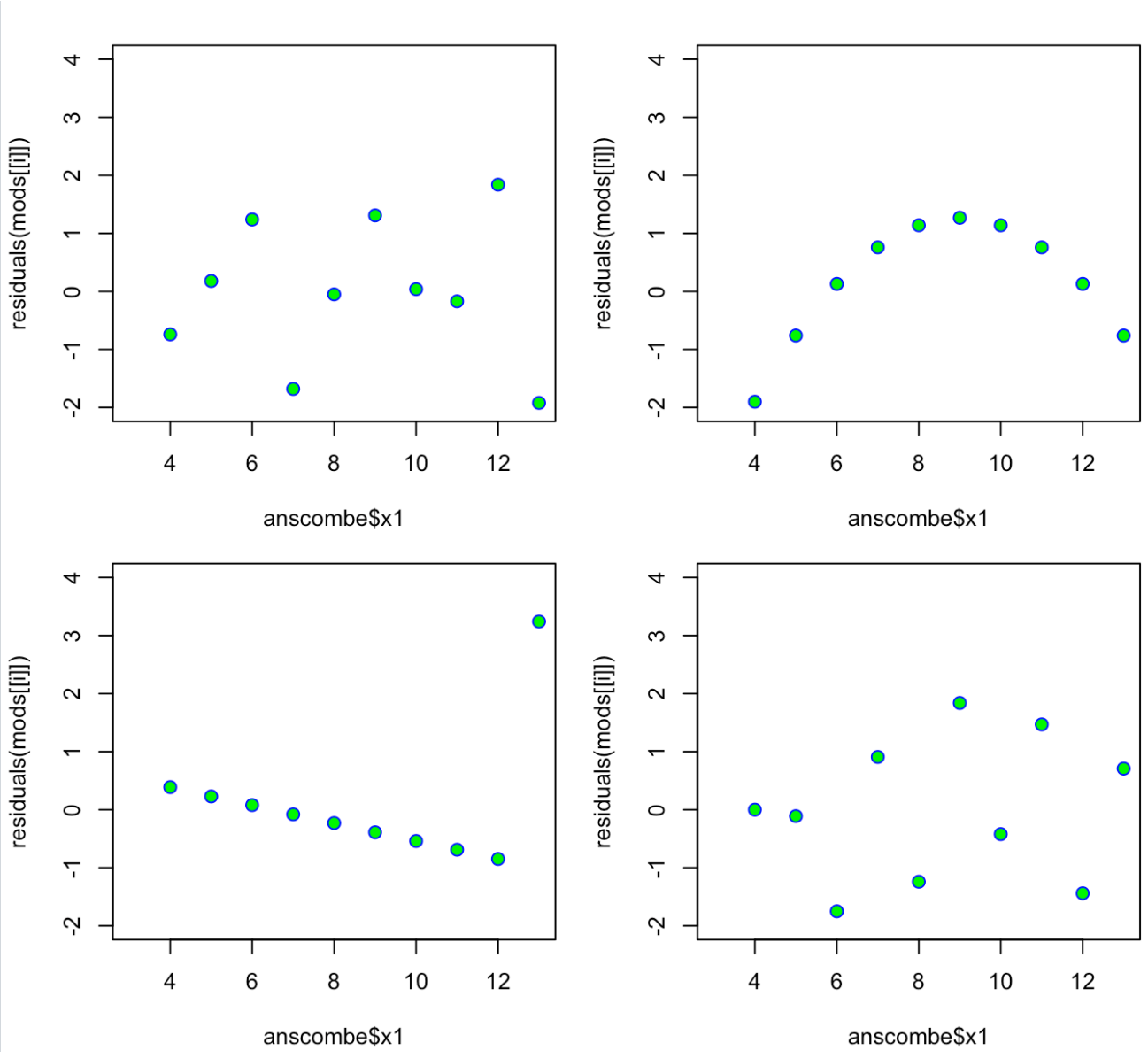
| depvar = y1 | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.0000909 | 1.1247468 | 2.667348 | 0.025734051 |
| x1 | 0.5000909 | 0.1179055 | 4.241455 | 0.002169629 |

| depvar = y2 | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.000909 | 1.1253024 | 2.666758 | 0.025758941 |
| x2 | 0.500000 | 0.1179637 | 4.238590 | 0.002178816 |

| depvar = y3 | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.0024545 | 1.1244812 | 2.670080 | 0.025619109 |
| x3 | 0.4997273 | 0.1178777 | 4.239372 | 0.002176305 |

| depvar = y4 | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.0017273 | 1.1239211 | 2.670763 | 0.025590425 |
| x4 | 0.4999091 | 0.1178189 | 4.243028 | 0.002164602 |

# Residual Plots of the Anscombe Regressions




*Note: plot at lower right has wrong explanatory variable.*

# Understanding the Anscombe Data Set II

- Of course, the Anscombe data sets are *constructed* to have very similar regression output. That is *not* important.
- Remember, statistics
  1. takes some data, and
  2. algorithmically summarizes it for
  3. human interpretation.

  Normally the data is too large for a human to visualize it from the raw numerical matrix.
  - It needs to be ordered by some useful index, or displayed graphically.

# Modern statistical methods, data mining, and big data

- The subject of statistics has changed dramatically in the last two decades.
  - Partly due to development of statistical theory.
  - Partly due to availability of new kinds of data
    * "big data" from sensor systems, POS (point of sale) data, and social networks
    * panel data from large scale surveys, mostly of household and consumer behavior (primarily in the U.S.)
  - Partly due to the diffusion of cheap powerful computers such as GPUs

# Bayesian statistics

- Bayesian statistics is named in honor of the Reverend Thomas Bayes (1701–1761), for whom Bayes' Law is also named.
- One of the key ideas of Bayesian statistics is that "probability is orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information."[a]
- *Bayes' Law*, $\mathrm{P}[A|B] = \mathrm{P}[A]\frac{\mathrm{P}[B|A]}{\mathrm{P}[B]}$, is the simplest case. It shows how to *update* the probability that $A$ occurs, given the information that $B$ occurred.

[a]Edwards, W., H. Lindman, and L. J. Savage [1963]. "Bayesian Statistical Inference for Psychological Research." *Psychological Review*. **70**: 193–242. doi:10.1037/h0044139, pp. 519–520.

# Using Bayes' Law: The taxi accident case

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue.The court tested the reliability of the witness under the circumstances that existed the night of the accident and concluded the witness identified each one of the colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved was Blue rather than Green?

# Using Bayes' Law: Defining the events

To use Bayes' Law, we need to define events A and B. They are

- A = the cab in the accident was blue
- B = the witness identified the cab as blue

From the description,

- We want to calculate Pr[A|B].
- We know Pr[A] = 0.15 (based on the fraction of Blue cabs).
- We know Pr[B|A] = 0.80 (based on testing reliability of witness).
- Bayes' Law says Pr[A|B] = P[B|A]*P[A]/P[B] = 0.80*0.15/P[B]. But what is P[B]?

# Using Bayes' Law: Computing P[B]

This is tricky.

1. A law of probability says that the P[B] = P[B∩A] + P[B∩¬A].
2. Using the definition of conditional probability, we have P[B] = P[B|A]*P[A] + P[B|¬A]*P[¬A] = 0.80*0.15 + 0.20*0.85 = 0.29.

We then put this into Bayes' Law to get P[A|B] = 0.80*0.15/0.29 = 0.41.

# The prior and posterior distributions

- In *Bayesian statistics*, whole distributions are updated in the same way. The distribution before collecting data and updating is called the *prior distribution*, and the distribution after updating is called the *posterior* distribution.
- The posterior distribution is used for prediction or hypothesis testing in the same way we use any estimated distribution in statistics.
- The prior distribution summarizes the information we have in advance. It may be *informative*, *weakly informative*, or *uninformative* (also, *diffuse*).

# Informativeness of prior distributions

- An informative prior distribution might be used in the case of a distribution with a history, such as the distribution of temperatures on June 1 at Tokyo Station for the period 1901–2000. We can check on whether climate change is occuring by looking at the posterior distribution for the mean temperature after updating based on the temperatures for 2001–2017.
  - This check is called *analysis of variance* (ANOVA) and the $\chi^2$ test is used.
- A weakly informative prior distribution is used to constrain the posterior distribution to "reasonable" values. For example, we could use a normal distribution with a mean of 20° and a standard deviation of 5°, implying that our prior probability of a mean temperature less than 0° or greater than 40° is less than 0.001%.
- An uninformative distribution is one which is chosen arbitrarily, to constrain the posterior as little as possible. Uniform distributions (with implausibly wide range) are frequently used. The *Jeffreys distribution* is not a single distribution, but rather a systematic method for constructing uninformative distributions.

# The Bernstein-von Mises Theorem

- Weakly informative and uninformative priors tend to give results similar to frequentist statistics.
- The *Bernstein-von Mises Theorem* gives a Central Limit Theorem for Bayesian statistics. As usual there is convergence to a normal distribution with a large enough sample.
- The strength of the theorem is that convergence is regardless of prior.
- The theorem is difficult to prove and from a mathematician's viewpoint somewhat fragile, but good enough for practical research.

# The Bayes Factor

- Bayesian statistics offers an alternative to the usual significance-based approach for hypothesis testing. In the first of two articles comparing the Bayesian approach to the frequentist approach, Goodman[a] argues that the $p$-value (equivalently, significance level) is not a very good way to evaluate a single research result. If you take the $p$-value seriously, you have to interpret it as the expected fraction of erroneous claims in the long run, but it doesn't tell you about the strength of evidence for the current result.
- In the second article[b], he advocates use of the *Bayes factor* instead:

$$\beta = \frac{\Pr[data|H_0]}{\Pr[data|H_a]}$$

- The *smaller* the Bayes factor, the *stronger* the case for rejecting $H_0$ is.

---

[a] Goodman, Steven N. [1999] "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy." *Ann Intern Med.* **130**:995-1004.

[b] Goodman, Steven N. [1999] "Toward evidence-based medical statistics. 2: The Bayes factor." *Ann Intern Med.* **130**:1005-13.

# Bayes' Law with the Bayes Factor

- Then Bayes' Law is posterior odds of $H_0 = \beta \times$ prior odds of $H_0$ (where if $\Pr[E] = p$, the *odds* of $E$ are $p/(1-p)$).
- While the "prior odds of $H_0$" depends on some judgment of the researcher, $\beta$ *does not*. $\beta$ depends only on the definitions of $H_0$ and $H_a$ and the data, and so can be considered *objective*. It doesn't measure the *strength of the case* for $H_a$ (that depends on the priors), it measures the *strength of the new evidence* provided by the analysis.
- As usual, choosing the alternative hypothesis $H_a$ requires some judgement, but here there is an obvious default choice: the parameter under the hypothesis is equal to the observed value, which is the hypothesis best supported by the data.
  - This gives the *minimum Bayes factor*, or the strongest evidence against the null hypothesis.
- Note that the Bayes factor *is* a likelihood ratio (many classical hypothesis tests are based on likelihood ratios), but it is computed differently from the classical (frequentist) approach.

# Computing the minimum Bayes factor

The *minimum Bayes factor* when the null hypothesis is $H_0 : \mu = 0$ when $z = 1$ is observed is the ratio of the orange segment at $z = 1$ (about 0.14) to the blue segment at $z = 1$ (about 0.40), or about 0.35.