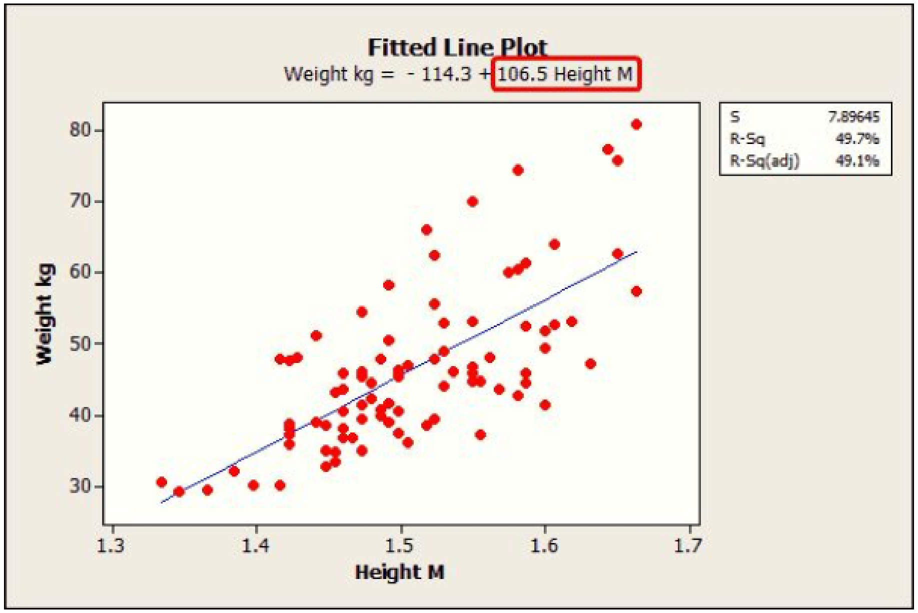
TODAY'S CLASS



1. Syllabus
2. Course Overview
3. Review of basic statistical principles

What is Statistics?

- *Statistics* is the science of collecting, organizing and drawing inferences from *data*
- Populations
- Samples
- Drawing *inferences* about the population, using *statistical* tools

Fundamental Concepts

- A *model* is a mathematical description of the quantities of interest
  - Example: Y has a normal distribution with unknown mean μ and standard deviation σ, often written $N(\mu, \sigma)$ or $N(\mu, \sigma^2)$ (need to distinguish which)

- A *parameter* is a numerical quantity that describes the population, usually unknown in practice
  - In the above example, μ and σ are the parameters

- A *statistic* is a value that we can calculate from a sample. It is often used to *estimate* a parameter, but it should not be confused with the parameter itself
  - The sample mean and the sample standard deviation

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i, \\ s_y^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2. \end{aligned}$$

Models

- There are many possible models
  - Gaussian
  - Binomial
  - Poisson
  - Gamma
  - Uniform

- The best known model of all is the Gaussian distribution, though you may know it by its other name: the *normal distribution*

Density functions (pdf's)

- Suppose we have random variable Y; we let the real number y be a generic value of the random variable Y, and we talk of its *density function* $f(y)$, also known as the *probability density function* and sometimes abbreviated *pdf*

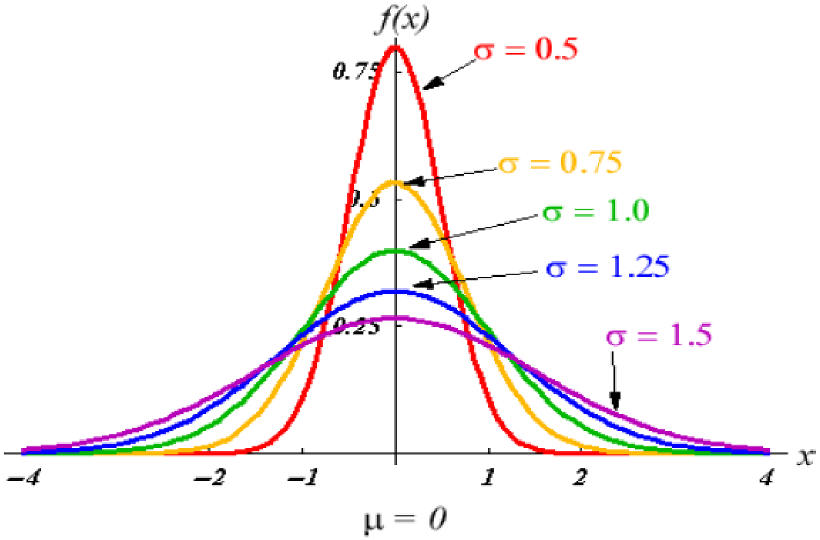
- Properties of $f(y)$:
  - $f(y) \geq 0$ for each y
  - The total area under the curve $f(y)$ is 1
  - For any a and b, the area under the curve between $y = a$ and $y = b$ represents the probability that Y is between a and b

- Example: for a normal distribution with mean μ and standard deviation σ,

$$f(y) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

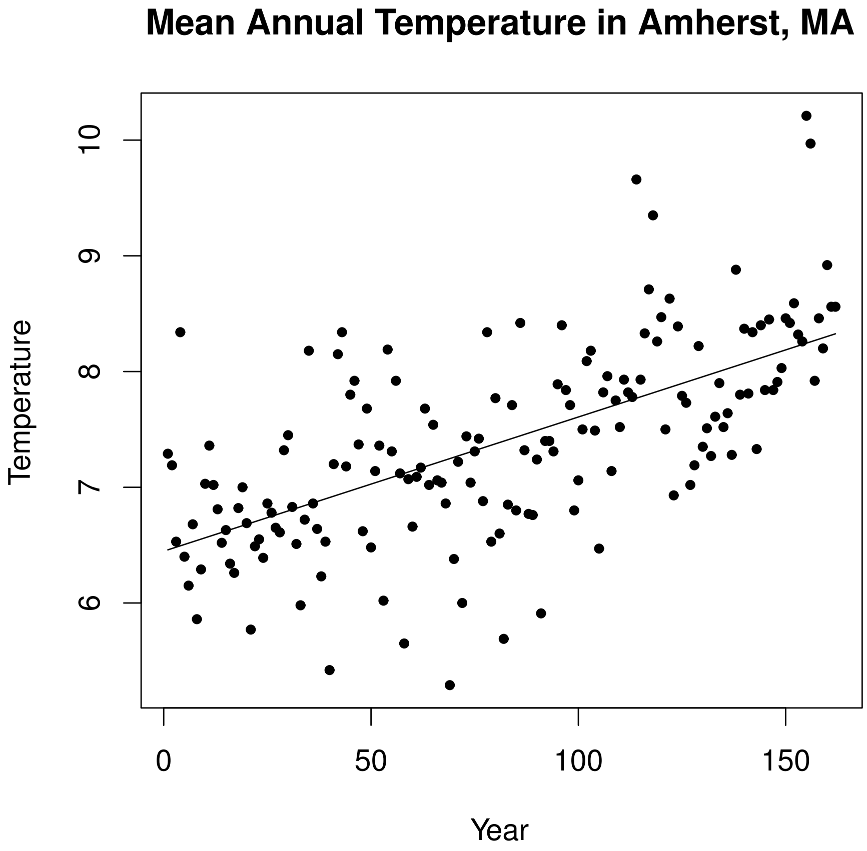
The Normal Distribution

Example: The normal distribution is the most important distribution in Statistics. Typical normal curves with different sigma (standard deviation) values are shown below.

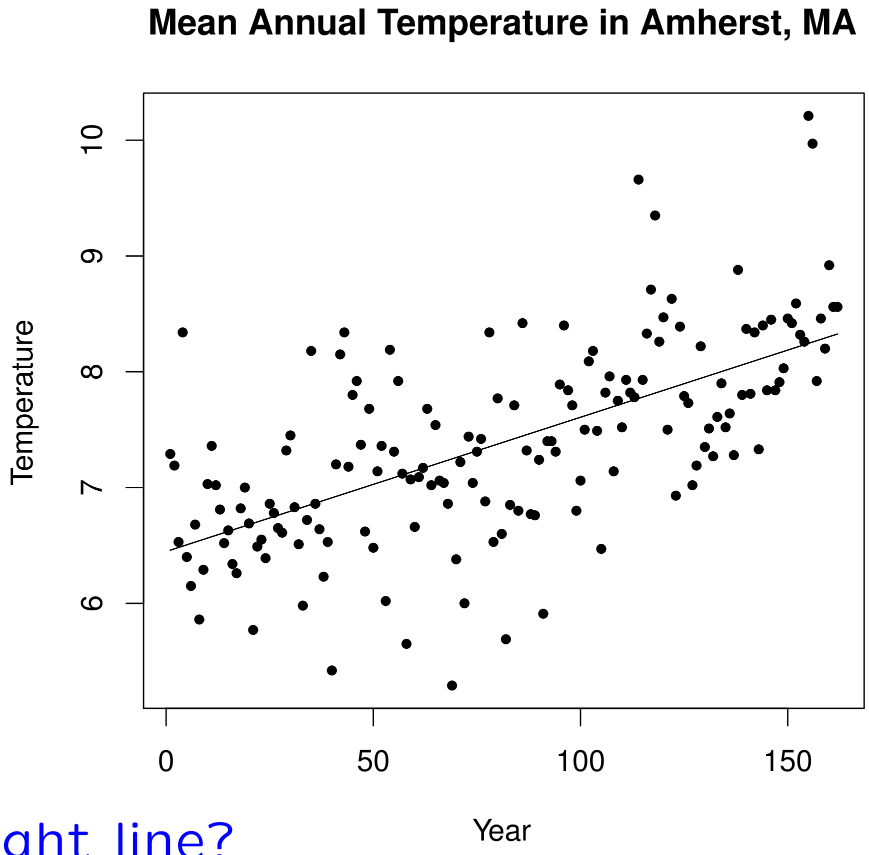


- Symmetric, unimodal, bell-shaped
- Completely specified by μ and σ
- The mean, median and mode are all the same

Example: Temperatures in Amherst, MA

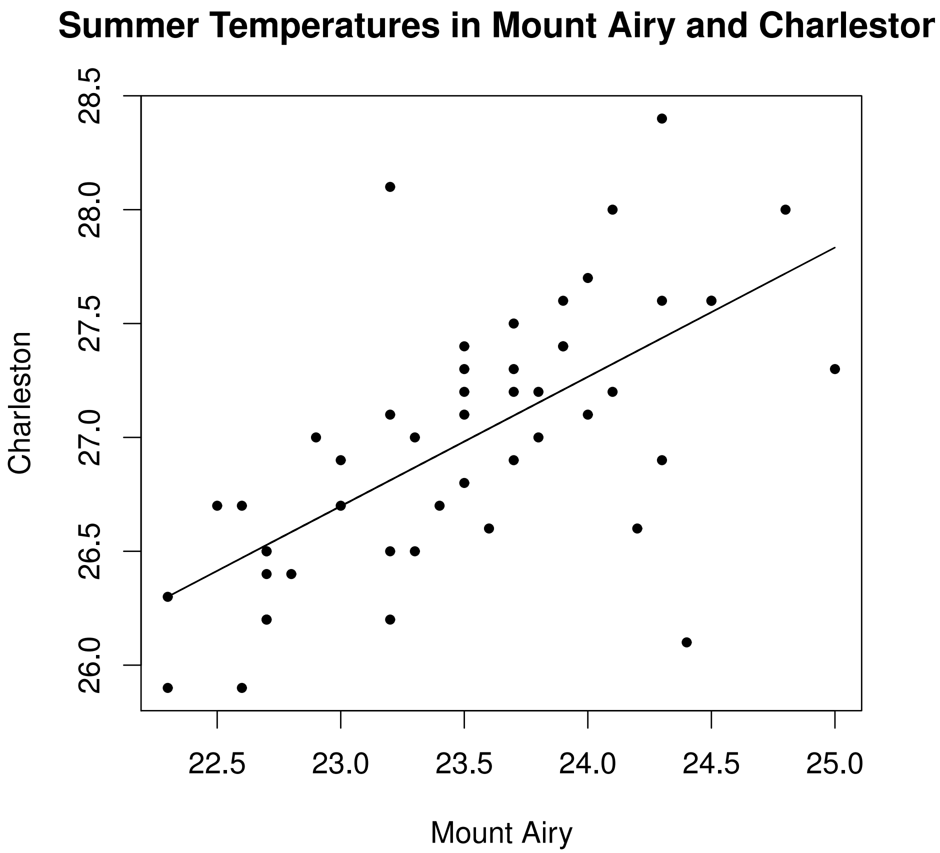


Example: Temperatures in Amherst, MA



- Is it a straight line?
- Are there outliers?
- Is the distribution approximately normal?

Temperatures in Mount Airy and Charleston



- How well can we predict the summer mean temperature in one city given the summer mean temperature in the other city?

Homework 1, due Monday, February 1
Very important: Show All Working!!

1. In a certain year, the mean SAT score for all students is 1200 and the standard deviation is 300. Assume that the distribution is normal.
   (a) What percentage of students scores above 1320? **[**3 points]
   (b) A certain college decides to give automatic acceptance to all students who score in the top 12% of all SAT scores. What SAT score does that correspond to? **[**3 points]

2. The file SATscores.csv in on the Data page in sakai. Using R, answer the following questions:
   (a) What are the sample mean $\bar{y}$ and the sample standard deviation s_y of this dataset? **[**2 points]
   (b) The *standard error* of a dataset is defined to be $s_y/\sqrt{n}$, where n is the sample size. For this dataset, what is the standard error? **[**2 points]
   (c) Draw a histogram of the data **[**3 points]
   (d) Draw a QQ-plot of the data **[**3 points]
   (e) Based on the histogram and the QQ-plot, would you say the data are normally distributed? **[**2 points]
   (f) A rough rule of thumb for when a sample mean is consistent with a hypothesized population mean (here, 1200) is that the difference between the two means should be less than 2 standard errors. Based on that, would you say the data are consistent with a population mean of 1200? **[**2 points]

Hints for Using R

1. Use the R functions `pnorm`, `qnorm`. For explanations, type `?pnorm` or `?qnorm`.
2. To read a csv file into R, type something like

   ```
   SAT=read.csv('SATscores.csv')
   ```

   You may need to insert the directory path before the file name.

   If you type `SAT` at the keyboard, you will get a listing of the data: the first few lines are

   ```
     Observation  SAT
   1           1 1190
   2           2 1240
   3           3 1120
   4           4 1430
   ```

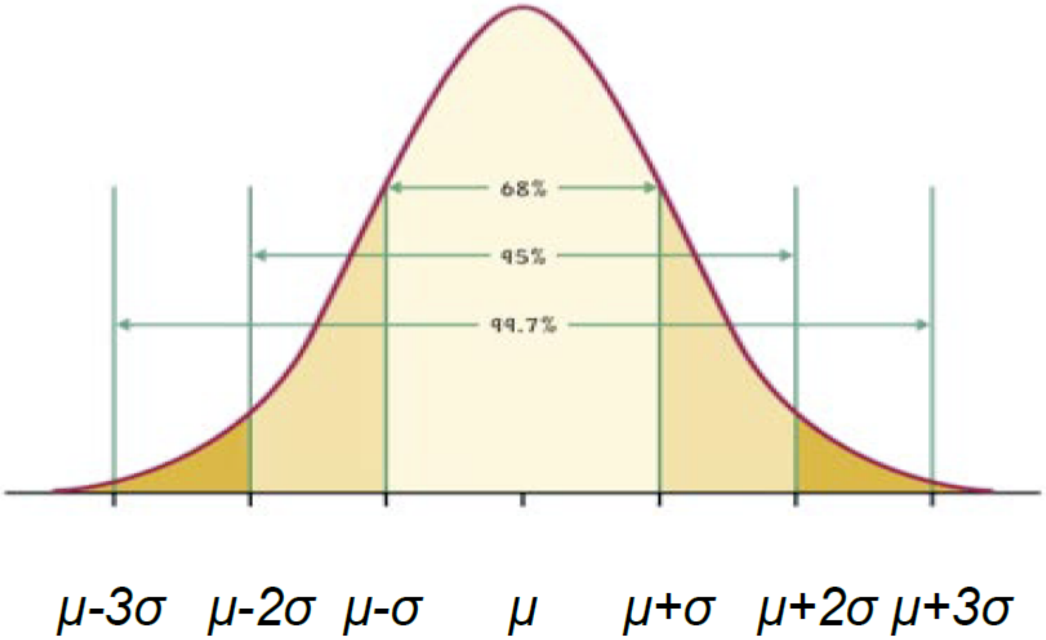
   The values in the second column are the ones you need. If you prefer a short variable name, say `y`, you can enter `y=SAT$SAT`.

   (a) The sample mean and variance of the vector `y` are given by `mean(y)` and `var(y)`. For standard deviation, `sy=sqrt(var(y))`.

   (b) The sample size is `length(y)`.

   (c,d) Use the commands `hist` and `qqnorm`. You can find more information by typing in `?hist` and `?qqnorm`.

The 68–95–99.7 Rule



- Approximately what percent of the distribution is 2 standard deviations above the mean?

- Between 1 and 3 standard deviations below the mean?

Typical Examples of a Normal Distribution

- Heights
- Weights
- Test scores (sometimes by design)
- Temperatures
- Rainfall??
- Incomes??
- Number of COVID cases on the UNC dashboard?
- These would all be *samples* from a larger *population*

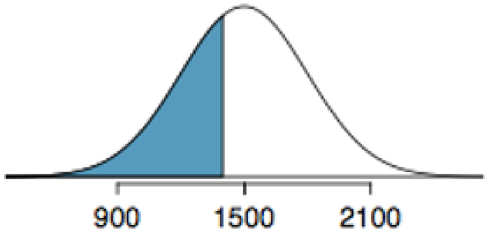
z-scores: Standardizing the normal distribution

- Suppose y is an observation from a normal distribution with mean μ and standard deviation σ
- The value $z = \frac{y-\mu}{\sigma}$ is call the *z-score*.
- If y indeed has a normal distribution with mean μ and standard deviation σ, then z has a normal distribution with mean 0 and standard deviation 1.
- Call *standard normal* — refer to standard tables or use software
- In R: function `pnorm(y,mu,sigma)` gives the left-hand tail probability of a normal distribution with mean mu and standard deviation sigma.
- `pnorm(y,mu,sigma)` is the same as `pnorm((y-mu)/sigma)` (try it!)

Example: SAT scores

- SAT scores from 2014 were normally distributed with mean 1500 and standard deviation 300

- What percentage of SAT scores were below 1400?



- The z-score is $\frac{1400-1500}{300} = -0.3333$

- `pnorm(-0.3333)= 0.3694539`, i.e. the answer is about 37%.

- Alternatively, `pnorm(1400,1500,300)= 0.3694413` (slight discrepancy due to rounding error)

Example: Quantiles/Percentiles

- SAT scores from 2014 were normally distributed with mean 1500 and standard deviation 300

- One university guarantees a scholarship to any student scoring in the top 3%.

- What score does that correspond to?

- On a standard normal, `qnorm(0.97)= 1.880794` (let's say 1.88 for simplicity)

- $z = \frac{y-\mu}{\sigma} = 1.88$, therefore $y = \mu + 1.88\sigma = 1500 + 1.88 * 300 = 2064$

- Alternatively, `qnorm(0.97,1500,300)=2064.238`.

- The required score is 2065 (to be conservative)

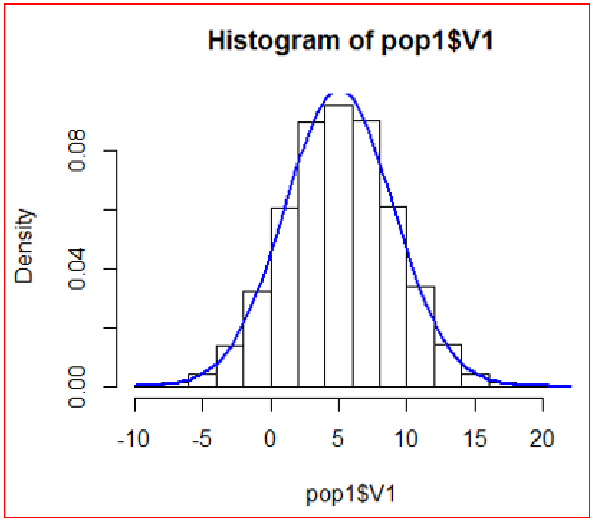
More Complicated Probabilities

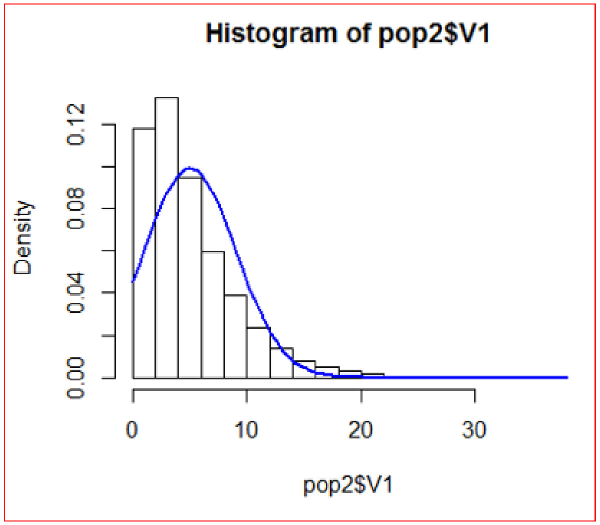
- What percentage of American males are between 68 and 72 inches tall?

- Google: $\mu = 69.2$, $\sigma = 2.66$ (inches)

- R: `pnorm(72,69.2,2.66)-pnorm(68,69.2,2.66)`=???

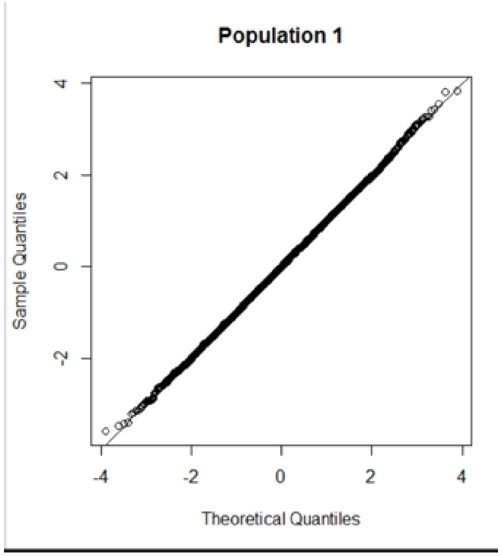
Histograms

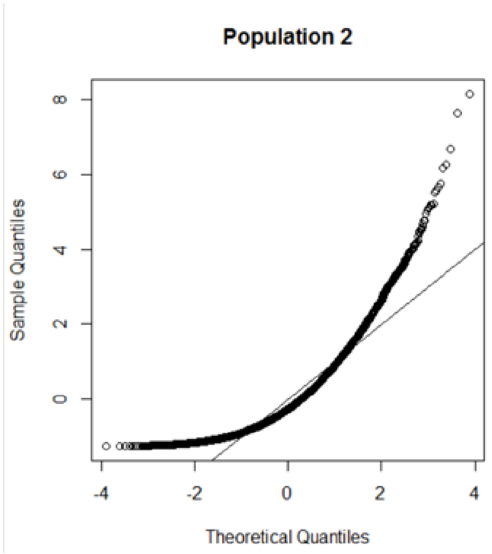




- “hist” in R
- Left plot: unimodal, symmetric, bell-shaped — consistent with normal distribution
- Right plot: clearly asymmetric, not a good candidate for normal distribution

QQ-Plots





- “qqnorm” in R
- Left plot: close to straight line, good fit to normal
- Right plot: strong curvature, probably not normal
- Not all cases are as clear-cut as these!

PARAMETERS

- A *parameter* is something that describes the *population* of interest.

- Assume a *finite population*, size N, values $Y_1, Y_2, \ldots, Y_N$.

- Common examples:
  - Mean, $\mu_Y = \frac{1}{N}\sum_{i=1}^{N} Y_i$
  - Standard Deviation, $\sigma_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu_Y)^2}$
  - Correlation coefficient between two variables X and Y, $\rho_{X,Y} = \frac{\sum_{i=1}^{N}(Y_i-\mu_Y)(X_i-\mu_X)}{\sqrt{\sum_{i=1}^{N}(Y_i-\mu_Y)^2\sum_{i=1}^{N}(X_i-\mu_X)^2}}$
  - Population proportion p, e.g. the fraction of the whole population that supports President Biden

MULTIVARIATE POPULATIONS

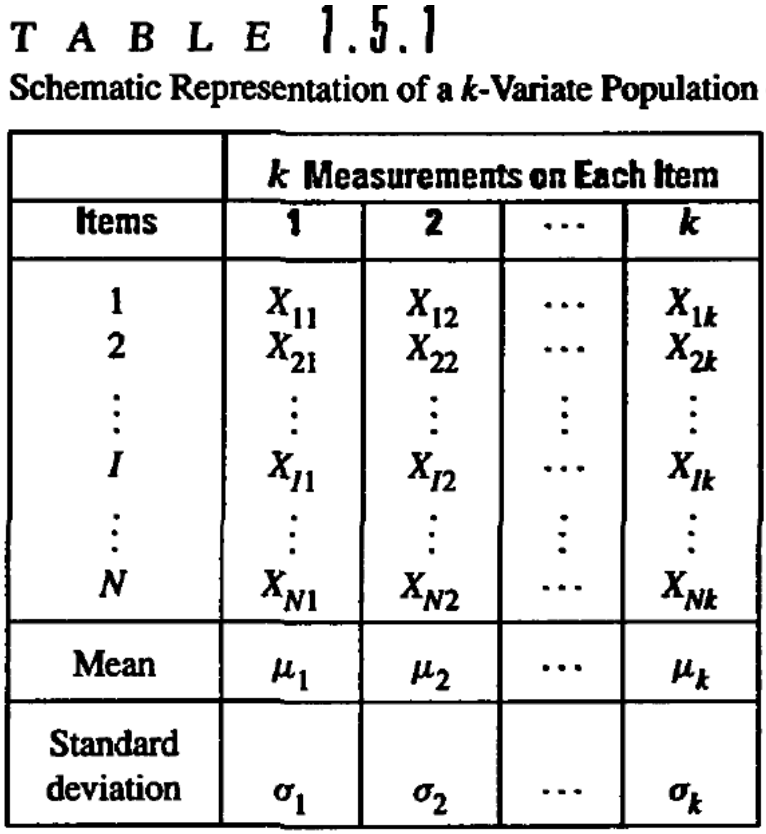


TABLE 1.5.1

Schematic Representation of a *k*-Variate Population of Size *N*

	k Measurements on Each Item			
Items	1	2	$\cdots$	*k*
1 2 $\vdots$ *I* $\vdots$ *N*	X_{11} X_{21} $\vdots$ X_{I1} $\vdots$ X_{N1}	X_{12} X_{22} $\vdots$ X_{I2} $\vdots$ X_{N2}	$\cdots$ $\cdots$ $\vdots$ $\cdots$ $\vdots$ $\cdots$	X_{1k} X_{2k} $\vdots$ X_{Ik} $\vdots$ X_{Nk}
Mean	μ_1	μ_2	$\cdots$	μ_k
Standard deviation	σ_1	σ_2	$\cdots$	σ_k

- k means, k standard deviations, $\frac{k(k-1)}{2}$ correlation coefficients

STATISTICAL INFERENCE

- Point Estimates
- Confidence Intervals (Interval Estimates)
- Hypothesis Tests

Point Estimates

- Assume a *sample* $y_1, ..., y_n$ from the *population* $Y_1, ..., Y_N$ (for correlation: also $x_1, ..., x_n$ from $X_1, ..., X_N$)

- Simple random sample (SRS): All $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ samples are equally likely

- Mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- Standard deviation: $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

- Correlation coefficient: $r_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

- These are all *unbiased* estimates in the following sense: the *mean of the estimate*, over many repetitions of the SRS, is equal to the population mean.

Example of Unbiased Estimation

R code `SDsimulation.txt`

```
SAT=read.csv(’C:/Users/rls/aug20/UNC/STOR455/Data/SATscores.csv’,header=T)
# population
Y=SAT$SAT
N=length(Y)
VAR=sum((Y-mean(Y))^2)/(N-1)
print(VAR)
# one sample of size 10
n=10
S=sample(1:N,n)
y=Y[S]
var=sum((y-mean(y))^2/(n-1))
# do this 100,000 times
varsum=0
for(i in 1:100000){
S=sample(1:N,n)
y=Y[S]
varsum=varsum+sum((y-mean(y))^2/(n-1))
}
print(varsum/100000)
```

Exercise: What if we divide by N and n instead of N-1 and n-1?

Quick Note about Unbiased Estimates

- The previous example illustrated the *variance* (square of the standard deviation) rather than the standard deviation itself

- Why was that?

- In fact, the sample standard deviation $\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$ is not exactly unbiased as an estimate of the population standard deviation (though it's close, and still better if you divide by $n-1$ instead of n)

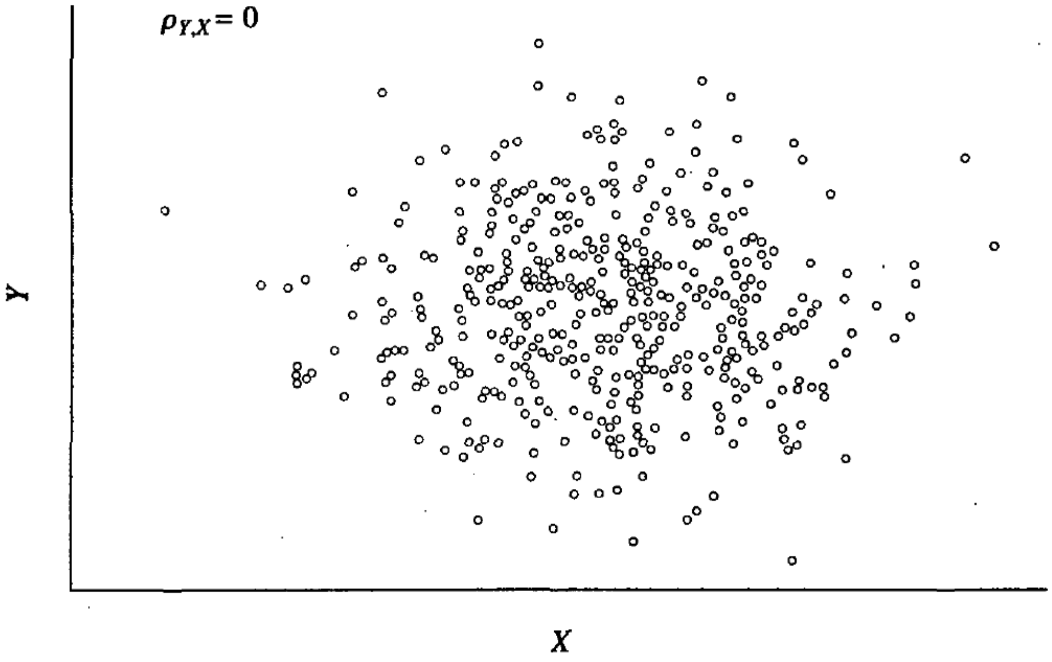
- This actually illustrates a difficulty about the concept of unbiasedness — it may seem like a natural and intuitive concept, but it's not so easy to achieve in practice

- ("Avoidance of bias" seems like a universal principle for statistics, but it's not quite that simple)

- Nevertheless, almost everyone divides by $n-1$ when estimating a standard deviation

Estimating Correlations I

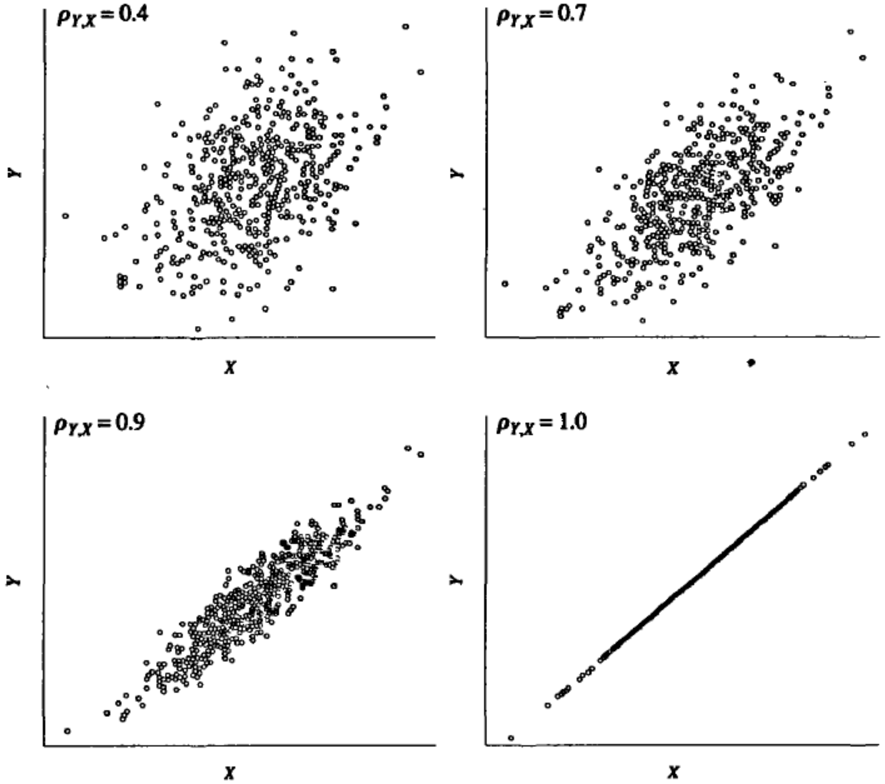
FIGURE 1.5.1



- Example of a scatterplot where the correlation is 0.

Estimating Correlations II

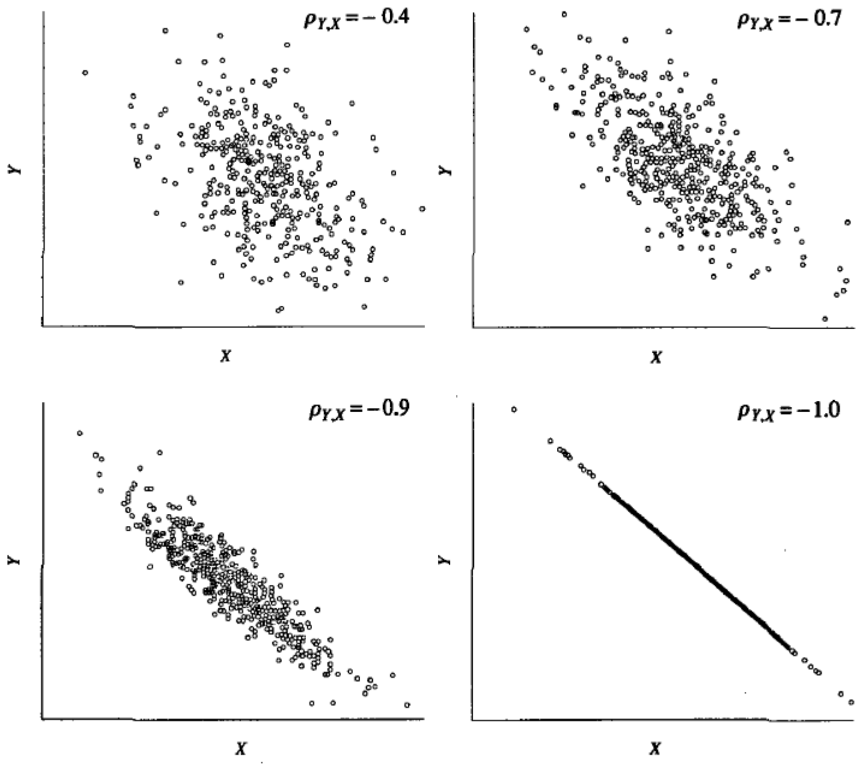
FIGURE 1.5.2



- Examples of scatterplots where the correlations are positive

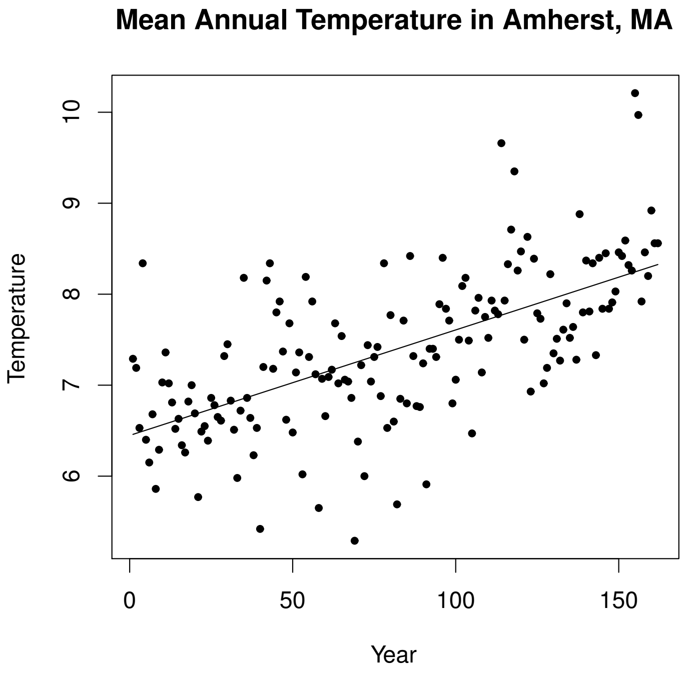
Estimating Correlations III

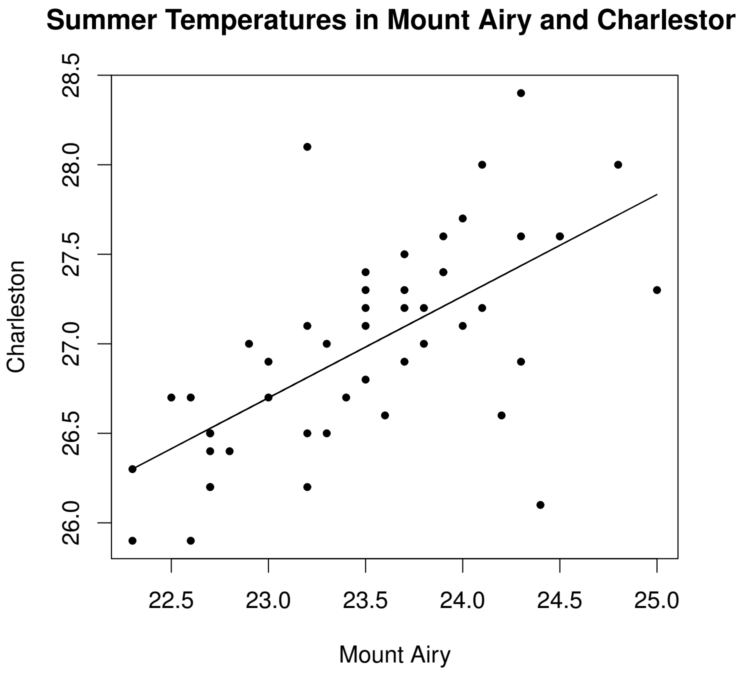
FIGURE 1.5.3



- Examples of scatterplots where the correlations are negative

Correlations in the Amherst and Mount Airy Datasets





Guess the correlation coefficients

Look at the estimates —

```
> cor(Amh)
          year      temp
year 1.0000000 0.6338541
temp 0.6338541 1.0000000
```

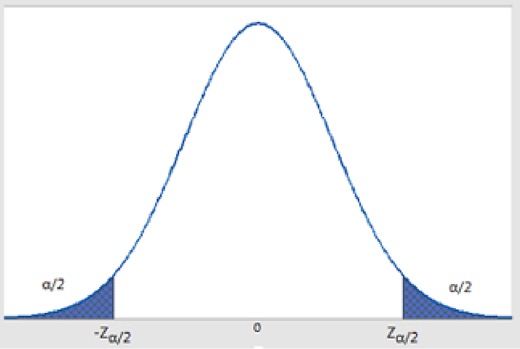
How would you interpret this table?

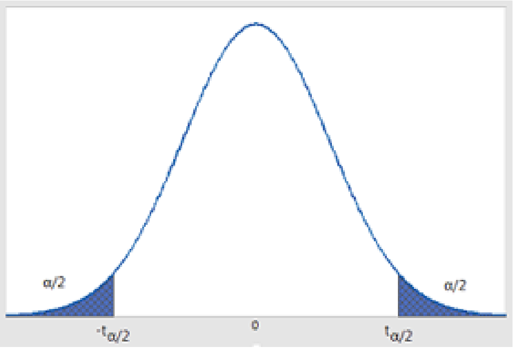
```
> cor(Mta)
                   Year       MtAiry Charleston
Year        1.000000000 -0.005637434 0.06387079
MtAiry     -0.005637434  1.000000000 0.65346598
Charleston  0.063870792  0.653465976 1.00000000
```

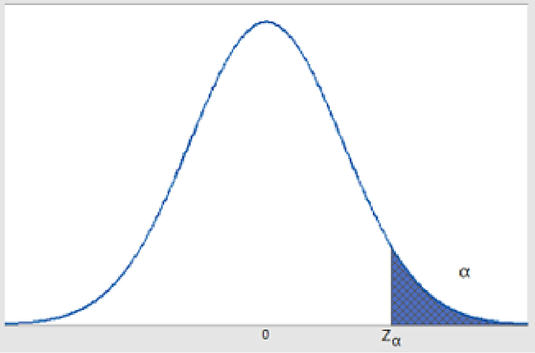
Confidence Intervals for the Mean

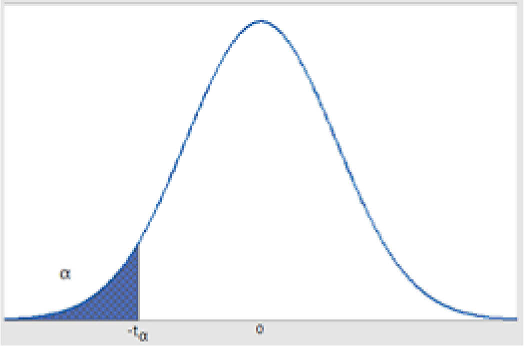
- Recall from STOR 155:
- $\bar{y} \pm z^* \frac{\sigma}{\sqrt{n}}$ if the SD σ is known
- $\bar{y} \pm t^* \frac{s_y}{\sqrt{n}}$ if σ is unknown and estimated by s_y
- z^* is derived from the normal distribution and t^* from the t distribution with $n-1$ degrees of freedom
- The values of z^* or t^* also depend on the *confidence coefficient* $1-\alpha$. Usually, $\alpha = 0.05$ for a 95% confidence interval, but that's not universal

Determining z^* or t^*









- One-sided or two-sided
- Confidence intervals are nearly always *two-sided* and *symmetric*
- One-sided bounds may be used in hypothesis testing

Theory of Confidence Intervals I

- Idea is to find *statistics* L and U (functions of the data) so that

$$\text{Pr}\left\{L \le \theta \le U\right\} \;=\; 1 - \alpha$$

where θ is the *parameter* of interest

- Case σ known: $Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution (mean 0, variance 1)

- Find $z^* = z_{1-\alpha/2}$ so that $\text{Pr}\left\{Z < -z_{\alpha/2}\right\} = \text{Pr}\left\{Z > z_{\alpha/2}\right\} = \alpha/2$.

- Then

$$1 - \alpha = \text{Pr}\left\{-z^* \le \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \le z^*\right\} = \text{Pr}\left\{\bar{y} - z^*\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{y} + z^*\frac{\sigma}{\sqrt{n}}\right\}$$

Theory of Confidence Intervals II

- Thus, $L = \bar{y} - z^* \frac{\sigma}{\sqrt{n}}$, $U = \bar{y} + z^* \frac{\sigma}{\sqrt{n}}$ fulfil the conditions for a confidence interval

- $\alpha = 0.05$: $z^* = 1.96$ (`qnorm(0.975)`), often rounded to 2 in practice (hence the rule of thumb for HW1, Q2(f))

- Case σ unknown: $t = \frac{\bar{y}-\mu}{s_y/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom, written t_{n-1}.

- Find $t^* = t_{n-1,1-\alpha/2}$ so that $\Pr\{t_{n-1} < -t^*\} = \Pr\{t_{n-1} > t^*\} = \alpha/2$. (In R: `qt(1-alpha/2,n-1)`.)

- In this case, $L = \bar{y} - t^* \frac{s_y}{\sqrt{n}}$, $U = \bar{y} + t^* \frac{s_y}{\sqrt{n}}$.

Prediction Intervals

- Suppose we are interested, not in *estimating* μ, but in *predicting* some future value $Y_0 \sim N[\mu, \sigma]$, where μ and σ are the same mean and standard deviation.

- We still use $\bar{y}$ as a point estimator/predictor, but now use the fact that $Y_0 - \bar{y}$ has mean 0 and variance $\sigma^2 \left(1 + \frac{1}{n}\right)$.

- $t = \frac{Y_0 - \bar{y}}{s_y \sqrt{1 + \frac{1}{n}}}$ has a t_{n-1} distribution

- The $100(1-\alpha)\%$ *prediction interval* for Y_0 is given by $\left(\bar{y} - t^* s_y \sqrt{1 + \frac{1}{n}}, \bar{y} + t^* s_y \sqrt{1 + \frac{1}{n}}\right)$ where $t^* = t_{n-1,1-\alpha/2}$, as before

Interval Estimation Summary

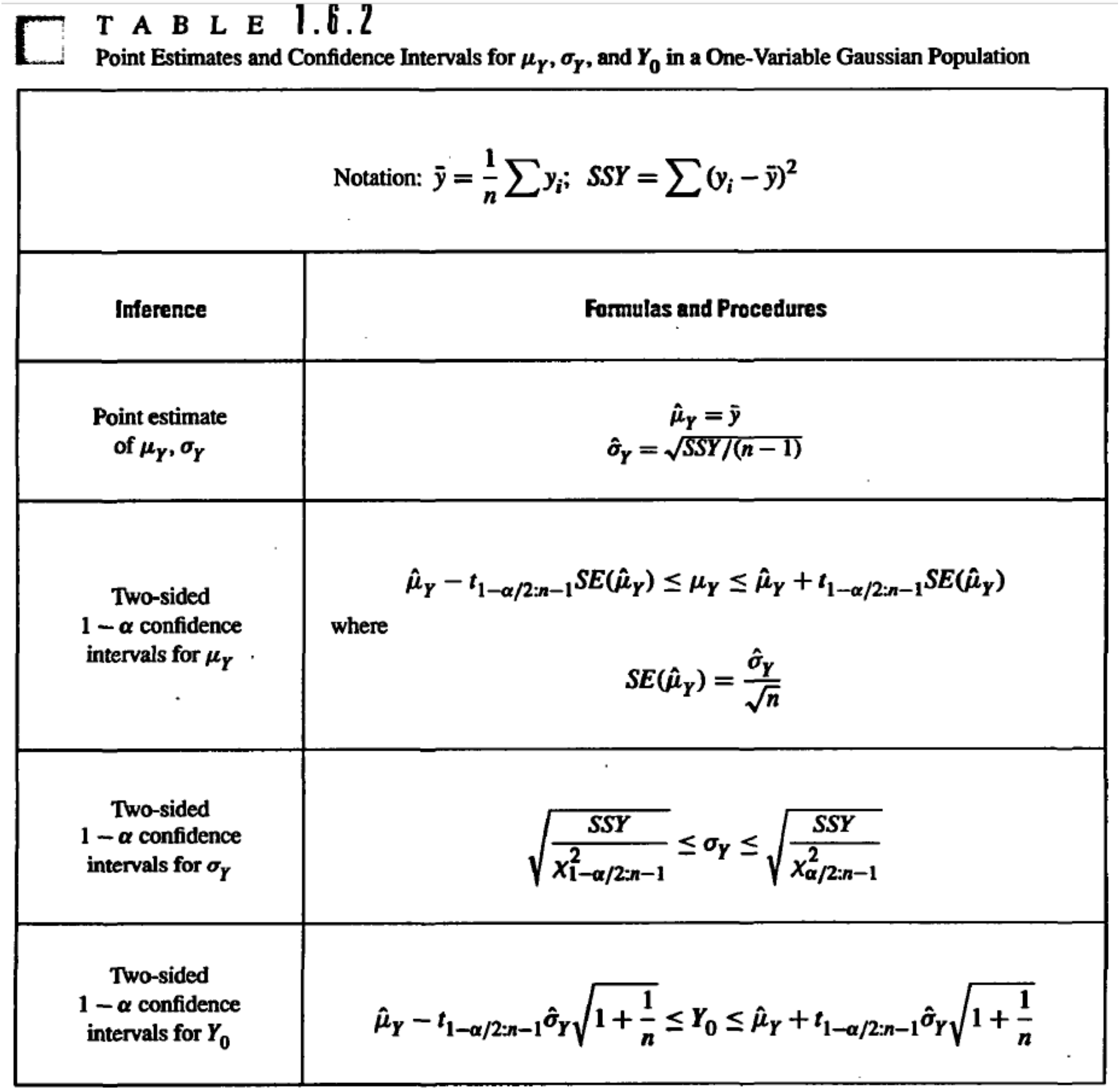


TABLE 1.6.2

Point Estimates and Confidence Intervals for μ_Y, σ_Y, and Y_0 in a One-Variable Gaussian Population

Notation: $\bar{y} = \frac{1}{n}\sum y_i$; $SSY = \sum (y_i - \bar{y})^2$

Inference	Formulas and Procedures
Point estimate of μ_Y, σ_Y	$\hat{\mu}_Y = \bar{y}$ $\hat{\sigma}_Y = \sqrt{SSY/(n-1)}$
Two-sided $1-\alpha$ confidence intervals for μ_Y	$\hat{\mu}_Y - t_{1-\alpha/2:n-1}SE(\hat{\mu}_Y) \le \mu_Y \le \hat{\mu}_Y + t_{1-\alpha/2:n-1}SE(\hat{\mu}_Y)$ where $SE(\hat{\mu}_Y) = \frac{\hat{\sigma}_Y}{\sqrt{n}}$
Two-sided $1-\alpha$ confidence intervals for σ_Y	$\sqrt{\frac{SSY}{\chi^2_{1-\alpha/2:n-1}}} \le \sigma_Y \le \sqrt{\frac{SSY}{\chi^2_{\alpha/2:n-1}}}$
Two-sided $1-\alpha$ confidence intervals for Y_0	$\hat{\mu}_Y - t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}} \le Y_0 \le \hat{\mu}_Y + t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}}$

Homework 2: Due Wednesday, February 10

- From the book: 1.6.1, 1.8.1, 1.8.3, 1.10.8, 1.10.9, 1.10.10, 1.10.11
  - Note: 1.6.1 requires the dataset `table164.txt`, on the Data page in sakai
- The dataset `EXAM.txt` (Data page in sakai) contains the midterm and final exam scores on a student exam.
  - Calculate a 95% confidence interval for (a) the mean score on the midterm, (b) the mean score on the final.
  - Test the hypothesis that the mean scores on the midterm and final are the same, against the alternative that they are different. Use $\alpha = 0.1$. What is your conclusion?
  - Draw a scatterplot of the final exam scores against the midterm exam scores, and draw a straight line through the plot (feel free to adapt the in-class examples for the Amherst and Mount Airy datasets). Would you say a straight line regression is justified in this case?

Instructions and Hints

- You may (and are expected to) use R for the computational part of any of this, but *show all working*: if you use R to get your answer, show the relevant R code so that we can see exactly how you got it, but also make sure that *you clearly and unambiguously state what your answer is*. You'd be amazed how many students neglect this very simple principle!

- Some of the Graybill-Iyer problems have solutions given in the "Answers" chapter of the book. I recommend that you work through these problems without first looking at the solutions, otherwise you won't learn much from trying to do them. However, I am not forbidding you to look at the solutions before handing them in: just make sure that the solution you hand in is *your* solution and contains a full explanation of what you did.

Homework 3: Due Wednesday, February 17

This is “Project 1.”

In sakai, go to “Resources”, then “Projects”, then “Project 1.pdf”. Full instructions are contained therein, but email the instructor on case of ambiguity.

Please note that the basketball data, to which the project refers, is also on sakai under “Data.”

Hypothesis Tests

- Interest in an unknown parameter θ (could be μ, or σ, or whatever...)
- *Null hypothesis* H_0, typically of the form $\theta = \theta_0$ for some specified value θ_0 (e.g. SAT example in HW1, θ was μ and the hypothesized value was $\theta_0 = 1200$)
- *Alternative hypothesis* H_1 or H_A, typically one of
  - $H_1 : \ \theta \neq \theta_0$
  - $H_1 : \ \theta > \theta_0$
  - $H_1 : \ \theta < \theta_0$
- Fix the *size* of the test (a.k.a. *significance level*) — the probability that we reject H_0, given that H_0 is true, must be no larger than α, for some suitable small value of α
- It is common to take $\alpha = 0.05$, but contrary to what some of my epidemiologist friends think, this is *not* a universal rule, e.g. could take $\alpha = 0.01$ or even 0.001 to get a more stringent rule
- Choose a *test statistic* T
- Define a *critical value* for T, usually written C
- The test will *reject* H_0 when $|T| > C$ or $T > C$ or $T < -C$, depending on the form of H_1

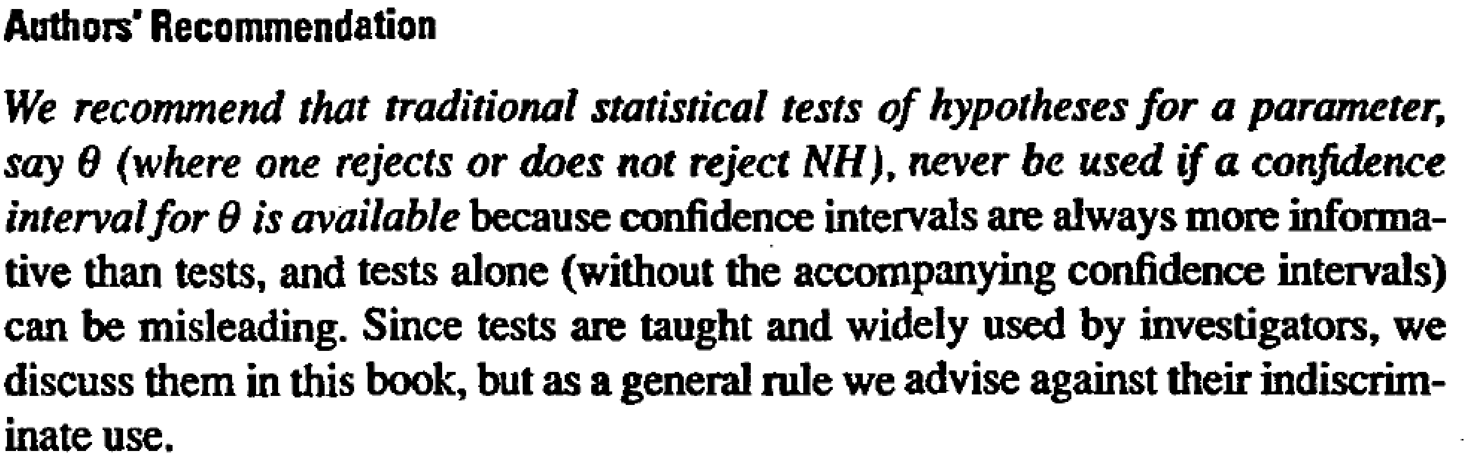
Example

- Sample of $n = 135$ SAT scores with mean $\bar{y} = 1227.5$, standard deviation $s_y = 147.7$.
- Test $H_0: \ \mu = \mu_0 = 1200$ versus $H_1: \ \mu \neq 1200$
- Test statistic $T = \frac{\bar{y}-\mu_0}{s_y/\sqrt{n}}$ with distribution t_{134} if H_0 is correct
- Choose $C = t_{134,0.975} = 1.978$ (`qt(0.975,134)` in R)
- Reject H_0 if $|T| > 1.978$.
- In fact $T = \frac{1227.5-1200}{147.7/\sqrt{135}} = 2.163$, so we reject H_0
- *Alternatively*, a 95% confidence interval for μ is $\bar{y} \pm t^* \frac{s_y}{\sqrt{n}} = 1227.5 \pm \frac{1.978 \times 147.7}{\sqrt{135}} = (1202.4, 1252.6)$
- How would this calculation change if H_1 was either $\mu > 1200$ or $\mu < 1200$?

One-sided Tests

- If $H_1 : \mu > 1200$:
  - Define $C = t_{n-1,1-\alpha} = t_{134,0.95} = 1.656$. We use $1 - \alpha = 0.95$ rather than $1 - \alpha/2 = 0.975$ (see figure on "Determining z^* or t^*" slide
  - $2.163 > 1.656$ so we again reject H_0
- If $H_1 : \mu < 1200$:
  - In this case we reject if $T < -C$
  - Here $2.163 > -1.656$ so we *accept* H_0
  - This might seem an odd conclusion given our previous results, but the practical conclusion is that $\mu = 1200$ is better supported by the data than any value for which $\mu < 1200$.

Tests or Confidence Intervals?



Authors' Recommendation

We recommend that traditional statistical tests of hypotheses for a parameter, say θ (where one rejects or does not reject NH), never be used if a confidence interval for θ is available because confidence intervals are always more informative than tests, and tests alone (without the accompanying confidence intervals) can be misleading. Since tests are taught and widely used by investigators, we discuss them in this book, but as a general rule we advise against their indiscriminate use.

My take?

- I tend to agree
- Many people (wrongly) interpret a decision to accept H_0 as proof that H_0 was correct
- This is not true — accepting H_0 often means only that there wasn't enough data to reject it
- A confidence interval is more informative because it gives you the full range of values of the unknown parameter that are consistent with the data
- Nevertheless, testing without a CI is still common practice...

p-values

- Another way of looking at hypothesis tests
- We observed a value $\bar{y} = 1227.5$. If H_0 is correct, $\mu = 1200$. How *improbable* is the value $\bar{y} = 1227.5$?
- If the data had a true normal distribution, the probability that $\bar{y} = 1227.5$ would be 0 (because it's a continuous distribution)
- In reality, the normal distribution isn't exact (SAT scores take integer values) but still, the probability that $\bar{y}$ is *exactly* 1227.5 is very small, regardless of the true value of μ
- A more meaningful question: what is the probability that $\bar{y}$ is *at least as extreme* as 1227.5, if $\mu = 1200$?
- $\text{Pr}\left\{\bar{y} \geq 1227.5\right\} = \text{Pr}\left\{\frac{\bar{y}-\mu}{s/\sqrt{n}} \geq \frac{1227.5-1200}{147.7/\sqrt{135}}\right\} = \text{Pr}\left\{T > 2.168\right\}$ where T is has a t_{134} distribution
- `pt(2.168,134,lower.tail=F)` gives the answer 0.016.

Interpretation

- 0.016 seems quite a small probability, but interpreting it is not so easy.
- If it was a *two-sided* test (i.e. if H_1 was $\mu \neq 1200$ rather than $\mu > 1200$), we should take into account that the deviation might have been equally far in the other direction ($\bar{y} = 1200 - 27.5 = 1172.5$)
- In this case, we should use a *two-sided p-value* — the distribution is symmetric around $\mu = 1200$, so

  $$\text{Pr}\,\{\bar{y} \geq 1227.5 \text{ or } \bar{y} \leq 1172.5\} \;=\; 2 \times 0.016 \;=\; 0.032.$$

  Still quite small (in particular, it's < 0.05), but not looking so dramatic
- The situation would be more complicated if, for example, we had tested several samples before finding one that was *statistically significant*. What would be the interpretation then?
  - Example: suppose I took data from Chapel Hill High, East Chapel Hill High, Carrboro High, Durham Academy and the NCSSM. In the first four, $\bar{y}$ is between 1180 and 1220, but at NCSSM, it's 1227.5. Is that *significant*?
- Problem of *simultaneous* or *multiple testing*

Simultaneous Confidence Intervals

- Suppose we have K populations, parameters $\theta_1, ..., \theta_K$.
- Would like to specify lower and upper confidence bounds L_k, U_k such that

$$\Pr\{L_k \leq \theta_k \leq U_k \text{ for each } k = 1, ..., K\} \geq 1 - \alpha. \tag{1}$$

- One way to achieve this is to set L_k, U_k so that

$$\Pr\{L_k \leq \theta_k \leq U_k\} = 1 - \frac{\alpha}{K} \text{ for } k = 1, ..., K. \tag{2}$$

- **Theorem:** If (2) holds, then so does (1).
- This is called *Bonferroni's Inequality*.

Example

- Five high schools, population mean SAT scores $\mu^{(1)}, ..., \mu^{(5)}$, sample means $\bar{y}^{(1)}, ..., \bar{y}^{(5)}$, sample SDs $s_y^{(1)}, ..., s_y^{(5)}$, sample sizes $n^{(1)}, ..., n^{(5)}$.
- Define $t^{(k)} = t_{n_k - 1, 1-\alpha/10}$ for $k = 1, ..., 5$.
- The confidence interval for $\mu^{(k)}$ is

$$\left(\bar{y}^{(k)} - \frac{t^{(k)} s_y^{(k)}}{\sqrt{n^{(k)}}}, \ \bar{y}^{(k)} + \frac{t^{(k)} s_y^{(k)}}{\sqrt{n^{(k)}}} \right)$$

- For example, if high school 5 is NCSSM with $\bar{y}^{(5)} = 1227.5$, $s_y^{(5)} = 147.7$, $n^{(5)} = 135$, and if we set $\alpha = 0.05$ as usual, we will define $t^{(5)} = t_{134,0.995} = 2.613$ (`qt(0.995,134)` in R).
- The confidence interval for NCSSM will be $\left(1227.5 - \frac{2.613 \times 147.7}{\sqrt{135}}, \ 1227.5 + \frac{2.613 \times 147.7}{\sqrt{135}}\right) = (1194.3, \ 1260.7)$.

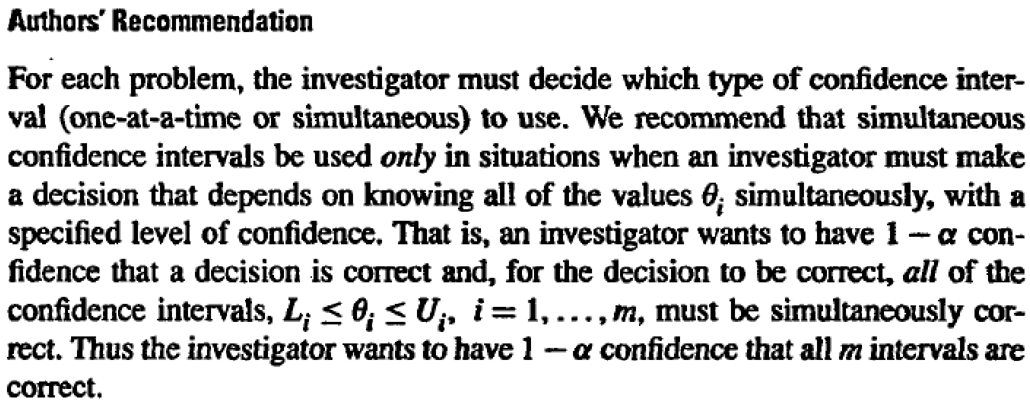
Multiple Testing

- Same idea ...
- Suppose we want to test K null hypotheses $H_0^{(1)}, ..., H_0^{(K)}$ against alternative hypotheses $H_1^{(1)}, ..., H_1^{(K)}$.
- Objective: define test statistics and critical regions so that

  $$\text{Pr}\left\{\text{Reject } H_0^{(1)} \text{ or Reject } H_0^{(2)} \text{ or ... or Reject } H_0^{(K)}\right\} \leq \alpha$$

  when each of $H_0^{(1)}, ..., H_0^{(K)}$ is true
- Bonferroni solution: set the significance level for each test to be α/K.
- There are other ways of constructing simultaneous confidence intervals or tests and whole books have been written about how to do it, but the Bonferroni method is the simplest and often almost as good as the other methods.

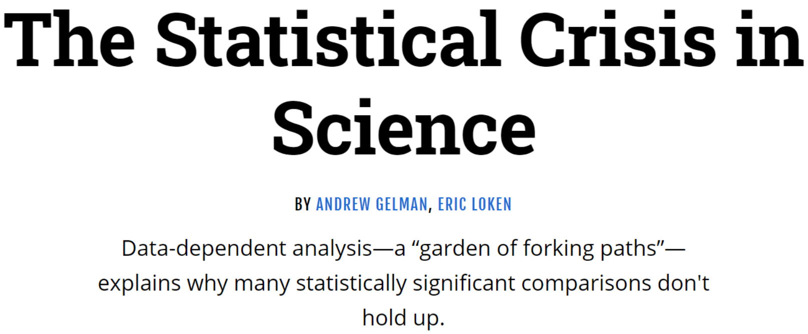
When to use these methods?



Authors' Recommendation

For each problem, the investigator must decide which type of confidence interval (one-at-a-time or simultaneous) to use. We recommend that simultaneous confidence intervals be used *only* in situations when an investigator must make a decision that depends on knowing all of the values θ_i simultaneously, with a specified level of confidence. That is, an investigator wants to have $1-\alpha$ confidence that a decision is correct and, for the decision to be correct, *all* of the confidence intervals, $L_i \le \theta_i \le U_i$, $i = 1, \ldots, m$, must be simultaneously correct. Thus the investigator wants to have $1-\alpha$ confidence that all m intervals are correct.

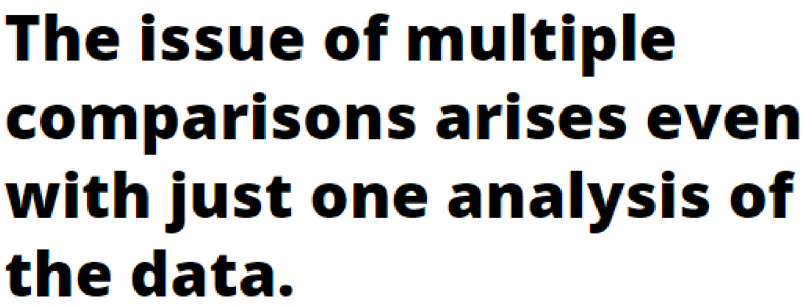
Maybe, but here's another view (Gelman-Loken paper in sakai)—



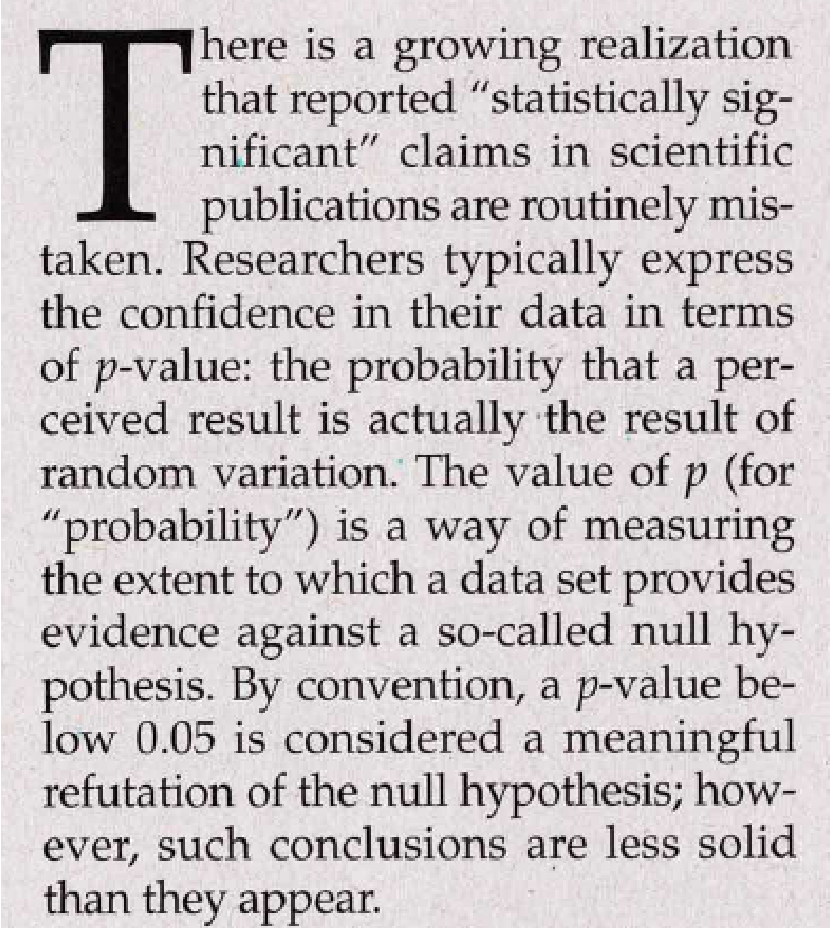
The Statistical Crisis in Science

BY ANDREW GELMAN, ERIC LOKEN

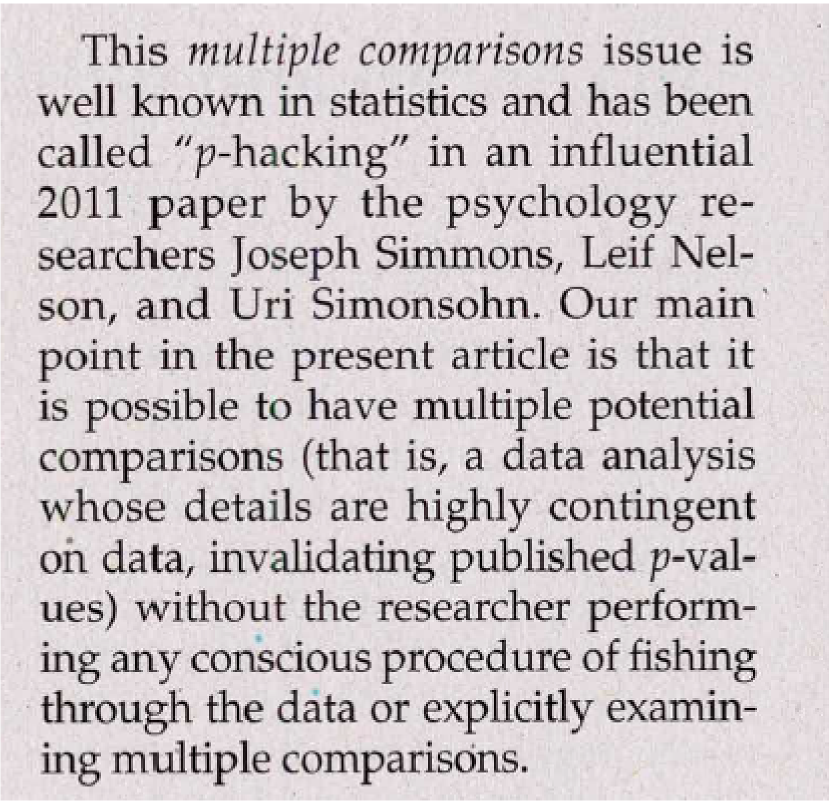
Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up.



The issue of multiple comparisons arises even with just one analysis of the data.



There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mistaken. Researchers typically express the confidence in their data in terms of *p*-value: the probability that a perceived result is actually the result of random variation. The value of *p* (for "probability") is a way of measuring the extent to which a data set provides evidence against a so-called null hypothesis. By convention, a *p*-value below 0.05 is considered a meaningful refutation of the null hypothesis; however, such conclusions are less solid than they appear.



This *multiple comparisons* issue is well known in statistics and has been called "*p*-hacking" in an influential 2011 paper by the psychology researchers Joseph Simmons, Leif Nelson, and Uri Simonsohn. Our main point in the present article is that it is possible to have multiple potential comparisons (that is, a data analysis whose details are highly contingent on data, invalidating published *p*-values) without the researcher performing any conscious procedure of fishing through the data or explicitly examining multiple comparisons.

Relevance to the current course

- You should be aware of the Bonferroni method, but it is not going to play a big role in this course (I don't think)

- There are other more specialized techniques (e.g. Scheffé and Tukey methods) that apply in specific situations — depending on how far I get, I may cover these are the end of the course

- Another solution to multiple comparison problems is simulation — another topic that I may include if there's time

- You should read the Gelman-Loken article as part of your general statistics education — the issues they raise are important, but there is no "magic bullet" to solving them.

Confidence Intervals and Hypothesis Tests for Comparing Two Means

- First sample: $x_1, ..., x_m$ from distribution $N[\mu_1, \sigma_1]$

- Second sample: $y_1, ..., y_n$ from distribution $N[\mu_2, \sigma_2]$

- Usually assume $\sigma_1 = \sigma_2 = \sigma$ (should do a rough check whether this is reasonable. but we won't do a formal test)

- Find a confidence interval for $\mu_1 - \mu_2$, or ...

- Test the null hypothesis $\mu_1 = \mu_2$ against any of
  - $\mu_1 \neq \mu_2$
  - $\mu_1 > \mu_2$
  - $\mu_1 < \mu_2$

Theory

- $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \sim N\left[\mu_1, \frac{\sigma}{\sqrt{m}}\right]$

- $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \sim N\left[\mu_2, \frac{\sigma}{\sqrt{n}}\right]$

- $\bar{x} - \bar{y} \sim N\left[\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}\right]$

- σ unknown: use the pooled estimate $s^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{m+n-2}$ with $m + n - 2$ degrees of freedom

- $\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$

Results

- Confidence interval for $\mu_1 - \mu_2$:

$$\bar{x} - \bar{y} \pm t^* s \sqrt{\frac{1}{m} + \frac{1}{n}}$$

- Hypothesis test: reject $H_0 : \ \mu_1 = \mu_2$ against $H_1 : \ \mu_1 \neq \mu_2$ if

$$|\bar{x} - \bar{y}| > t^* s \sqrt{\frac{1}{m} + \frac{1}{n}}$$

- In both cases, for significance level α, $t^* = t_{m+n-2,1-\alpha/2}$

- Question: for a one-sided test, how would you modify this?

Extensions

- What to do if $\sigma_1 \neq \sigma_2$?
  - No clear-cut solution, but there's an approximate solution called the *Welch-Satterthwaite formula* which you may have seen in STOR 155

- Extension to more than two samples: we could have K samples with means $\mu_1, \ldots \mu_K$ standard deviations $\sigma_1, \ldots \sigma_K$, test

  $$H_0 : \ \mu_1 = \ldots = \mu_K$$

  against

  $$H_1 : \ \mu_1, \ldots \mu_K \text{ are not all equal.}$$

- Again, we usually assume $\sigma_1 = \ldots = \sigma_K$ (but check this is reasonable)

- The procedure is then called "one-way analysis of variance" (`aov` function in R)

- We may talk about this later.

Other topics from Chapter 1 of the text

- Section 1.7 — functional notation
- Section 1.8 — vectors and matrices
- Section 1.9 — multivariate normal distribution

Section 1.7: Functional Notation

- Get used to expressions like

  $$f(x_1, x_2, \ldots, x_n)$$

  to denote a (scalar) function of n variables $x_1, \ldots, x_n$

- A function is *linear* if it can be expressed in the form

  $$f(x_1, x_2, \ldots, x_n) \;=\; a_0 + a_1 x_1 + \ldots + a_n x_n$$

  for some constants $a_0, a_1, \ldots, a_n$.

- *Example* $f(x_1, x_2, x_3) = 3x_1 - 7x_2 + 4x_3 + 19$ is linear, but $f(x_1, x_2, x_3) = 3x_1 - e^{x_2} + 4x_1 x_3 + 19$ is not.

Section 1.8: Vectors and Matrices

- A *matrix* is a two-dimensional array of numbers
- An $m \times n$ matrix has m rows and n columns
- Example: $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ is a 2×3 matrix
- An $n \times 1$ matrix is called a *column vector*, and a $1 \times n$ matrix is called a *row vector*. Usually, when we say "vector" without specifying whether it means a row vector or a column vector, we mean the latter.
- The *transpose* of a matrix, usually written A^T or A', is obtained by interchanging the rows and columns.
- With A as above, $A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$.

Operations on Matrices

- *Addition, Subtraction:* defined element by element. If A and B do not have the same dimensions, $A + B$ and $A - B$ are not defined.

- Example: if $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}$ then

$$A + B = \begin{pmatrix} 2 & 4 & 4 \\ 6 & 6 & 8 \end{pmatrix}, \quad A - B = \begin{pmatrix} 0 & 0 & 2 \\ 2 & 4 & 4 \end{pmatrix}.$$

- *Equality of Matrices:* Two matrices A and B are defined to be *equal* if and only if they are the same dimensions and every entry of A is equal to the corresponding entry of B.

- Example: if $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ then $A = B$.

Matrix Multiplication

- Suppose A is a $m\times n$ matrix, B is a $n\times p$ matrix, then $C = AB$ is defined, a $m \times p$ matrix, and its (i,j) entry is given by

  $$c_{i,j} \quad = \quad \sum_{k=1}^{n} a_{i,k}b_{k,j}$$

  (Sometimes the indexes are separated by commas, sometimes they are not, it really doesn't matter but try to be consistent)

- Example: If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$, then

  $$C \quad = \quad \begin{pmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 51 \end{pmatrix}.$$

Special Matrices

- The *zero matrix*, sometimes written $\mathbf{0}$, is an $m \times n$ array of zeroes, for example $A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$.

- A matrix is *square* if the row and column dimensions are the same, i.e. $m = n$.

- The $n \times n$ *identity matrix*, sometimes written I_n, is an $n \times n$ matrix with ones on the diagonal, zeroes elsewhere. For example, $I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

- A *diagonal* matrix is a square matrix whose only non-zero entries are on the diagonal, for example, $A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -15 \end{pmatrix}$.

Matrix Inverse

- If C is an $n \times n$ matrix, then the *inverse* of C, denoted C^{-1}, if it exists, is another $n \times n$ matrix with the property that

$$C^{-1}C \quad = \quad CC^{-1} \quad = \quad I_n.$$

- *Example*. If I slightly change my previous C to $C = \begin{pmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 52 \end{pmatrix}$, and define $D = \frac{1}{6}\begin{pmatrix} 32 & -24 & 6 \\ -31 & 33 & -12 \\ 6 & -12 & 6 \end{pmatrix}$, then $D = C^{-1}$.
- This is not so easy to derive (except in the 2 × 2 case, see next slide) but we shall see how to compute matrix inverses in R, and you can check directly by multiplying out that CD and DC are both the 3 × 3 identity matrix.
- Question: My original C was $\begin{pmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 51 \end{pmatrix}$ (different in $c_{3,3}$). Why did I change it?

Inverse of a 2×2 matrix

- If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

provided $ad \neq bc$.

Determinant of a Matrix

- If A is an $n \times n$ matrix, then there is a special quantity called the *determinant*, which is useful in calculating inverses.

- 2×2 case: if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $|A| = ad - bc$.

- Higher dimensions: not easy to calculate by hand, but the R function **`det`** does what you think.

- Most important property, if $|A| = 0$, then A^{-1} does not exist. It's the matrix equivalent of trying to divide by 0.

- However, there is actually something called a *generalized inverse*, which does something similar to A^{-1} when $|A| = 0$. The most famous form of generalized inverse is the *Moore-Penrose* inverse (e.g. Moore 1920, Penrose 1955).

- *Scientific trivia question:* who is Penrose, and why is he famous?

Implementation in R

- R is an *object-oriented* language — define matrices (and other mathematical objects) directly with R commands and to manipulate them according to the rules of matrix algebra

- To define a matrix, e.g.

  ```
  A=matrix(1:6,ncol=2,byrow=T)
  B=matrix(1:6,ncol=2,byrow=F)
  # note distinction between byrow=T and byrow=F: default is F
  t(B) # transpose
  # matrix operations: A+B, A-B are what you expect but
  # A*B is element by element multiplication
  C=A %*% t(B) # this is how you do matrix multiplication in R
  solve(C) # inverse of C
  det(C) # determinant of C
  ```

- Generalized inverse: `library(MASS)` and `ginv(..)`.

Section 1.9: Multivariate Gaussian Distributions

- Recall that a random variable Y is said to have a *Gaussian distribution* with mean μ and standard deviation σ if its probability density function (pdf) is $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$. In practice, verify this graphically (histograms, QQ plots)
- The vector $\begin{pmatrix} Y_1 & Y_2 & \dots & Y_K \end{pmatrix}^T$ is said to have a *multivariate Gaussian distribution* if, for any constants $a_1, \ a_2, \ \dots, \ a_K$, $\sum_{k=1}^{K} a_k Y_k$ has a univariate Gaussian distribution
- If a distribution is multivariate Gaussian, then it is characterized by the means $\mu_1, \dots \mu_K$, the standard deviations $\sigma_1, \dots \sigma_K$ and the correlations $\rho_{j,k}, \ 1 \le j, k \le K$.
- In reality, it's almost impossible to *prove* that a distribution is multivariate Gaussian, though sometimes we can find $a_1, \dots, a_K$ that prove it is not (see text for an example)
- Sometimes, it's easier to prove theoretically, e.g. if each of $Y_1, ..., Y_K$ is some linear combination of the same set of independent Gaussian $Z_1, \dots, Z_m$, then automatically, $Y_1, ..., Y_K$ are multivariate Gaussian

END OF CHAPTER 1!