Lecture 12: Beer and brawls and general bad assedness

Last time

We looked at a particular example of how to use the bootstrap to examine the represent information a random sample provides about a population parameter -- Hopefully we are recognizing the bootstrap as a useful device

At that point, we considered the relative risk of having a heart attack under Vioxx versus Aleve -- We constructed confidence intervals from the data for a 2000 study and put it in context of other studies happening at about the same time

Today

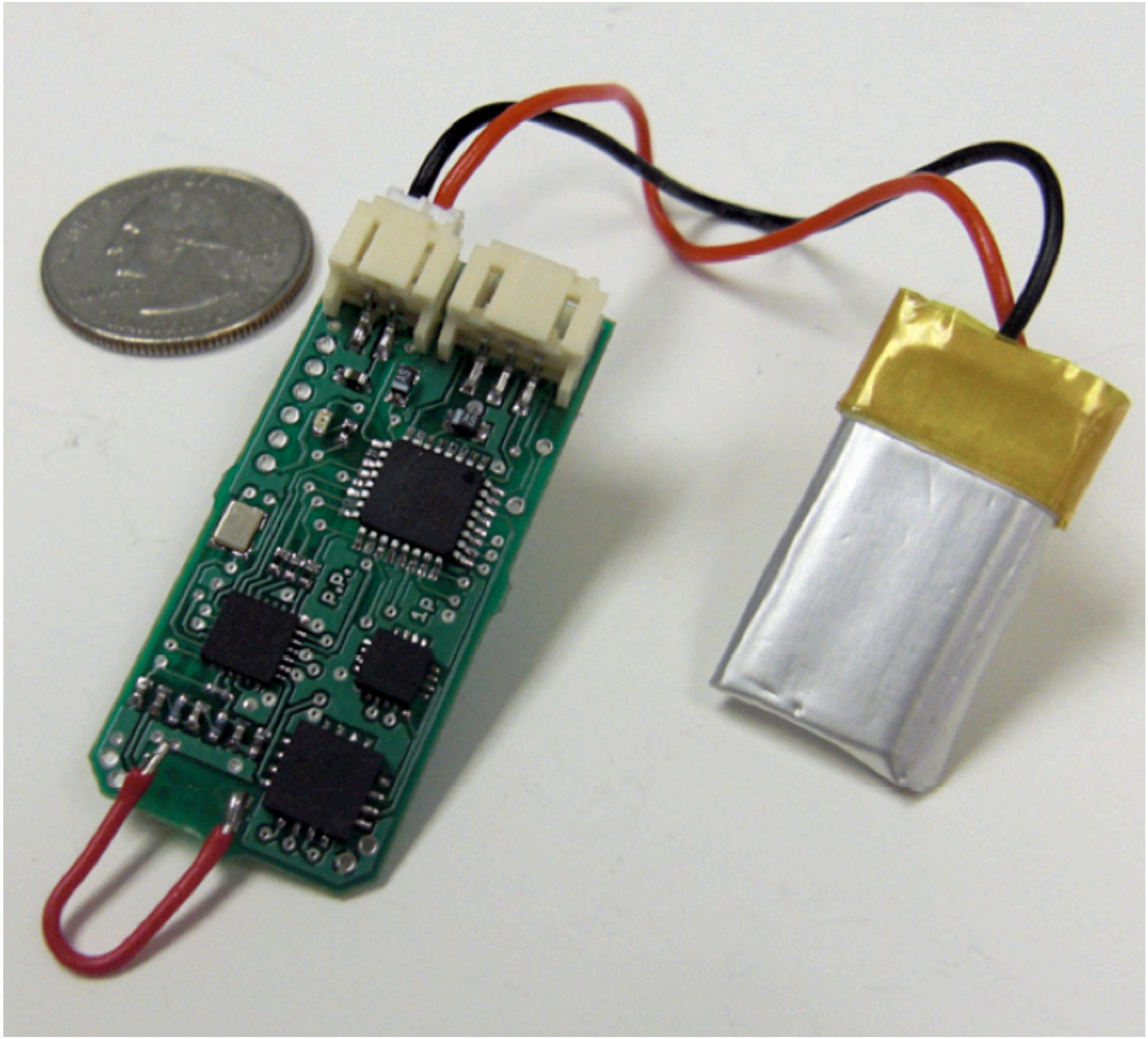
We are going to look at another way to derive the properties of the sampling distribution -- This approach relies heavily on assumptions about the data and is more limited than our bootstrap procedure

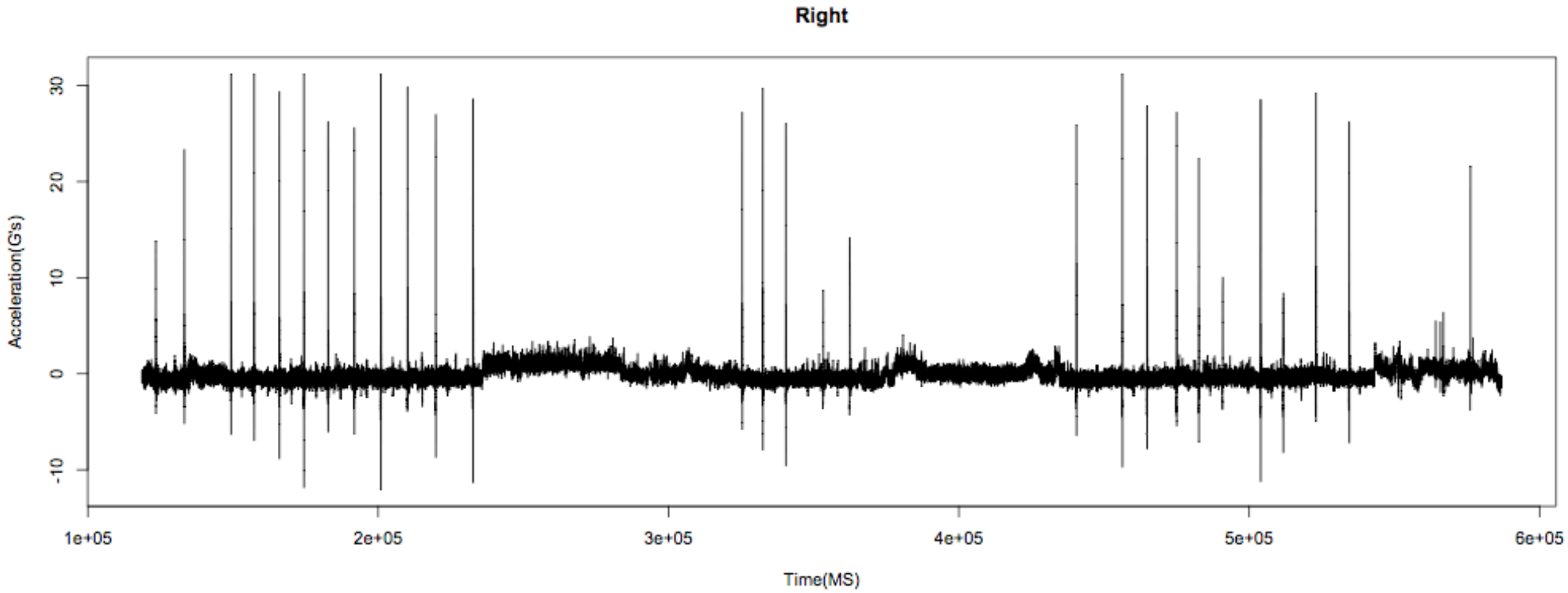
Thankfully, we'll see that when the assumptions of the mathematical approach hold true, the math solution and the bootstrap solution agree, making the bootstrap applicable in a wider range of cases

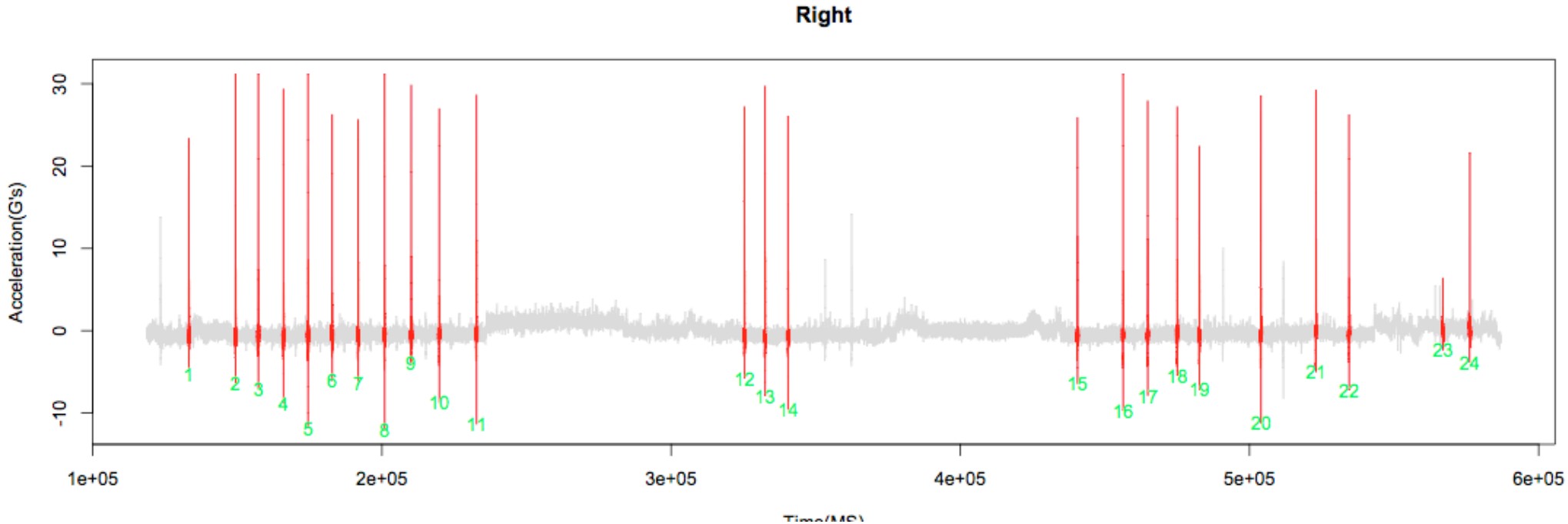
So today will be a little historical -- We're going to start however, in Gleason's Gym in Brooklyn...

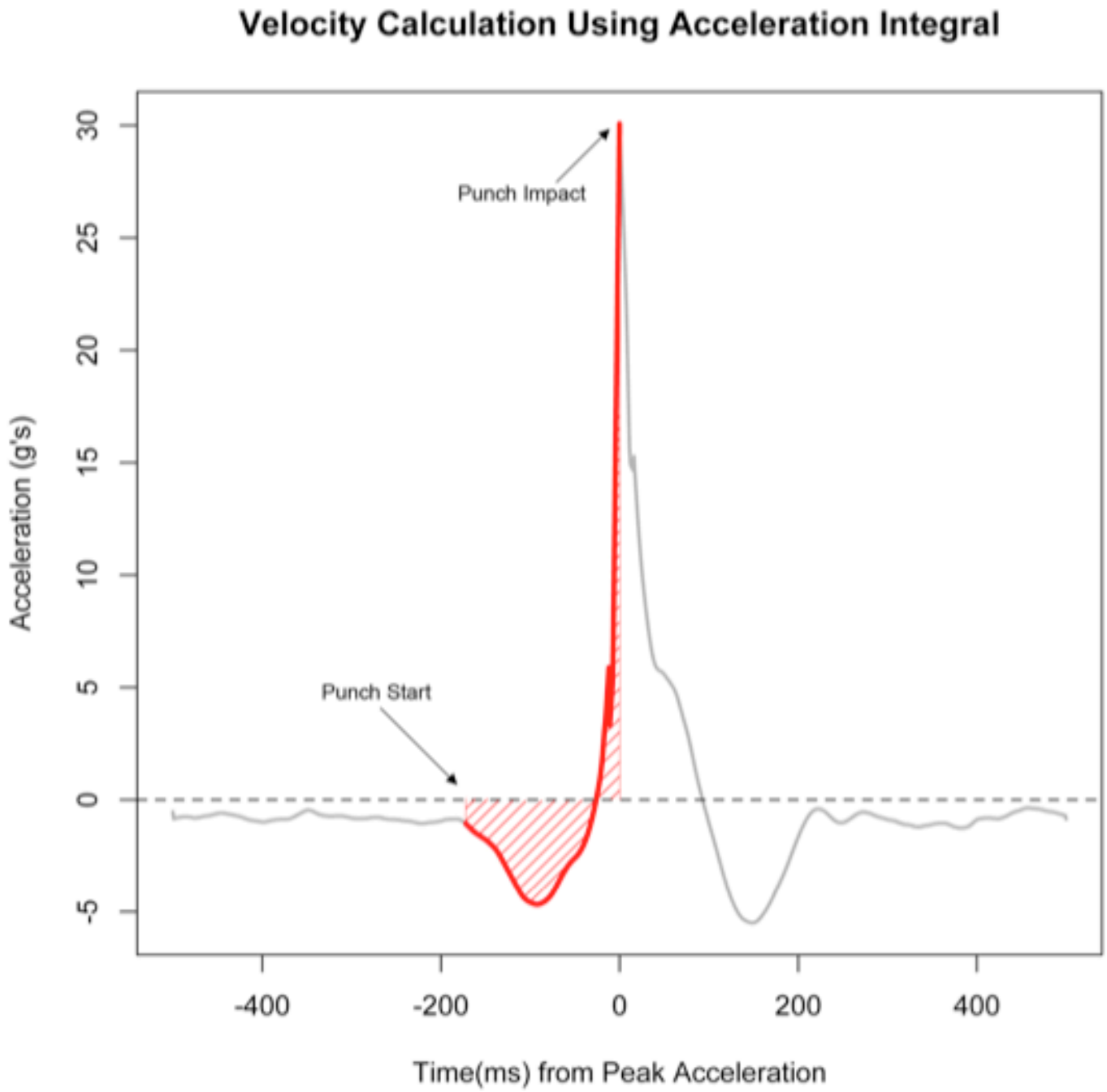

EXIT
SUPER
1970


WBO
GUZMAN
Reasons

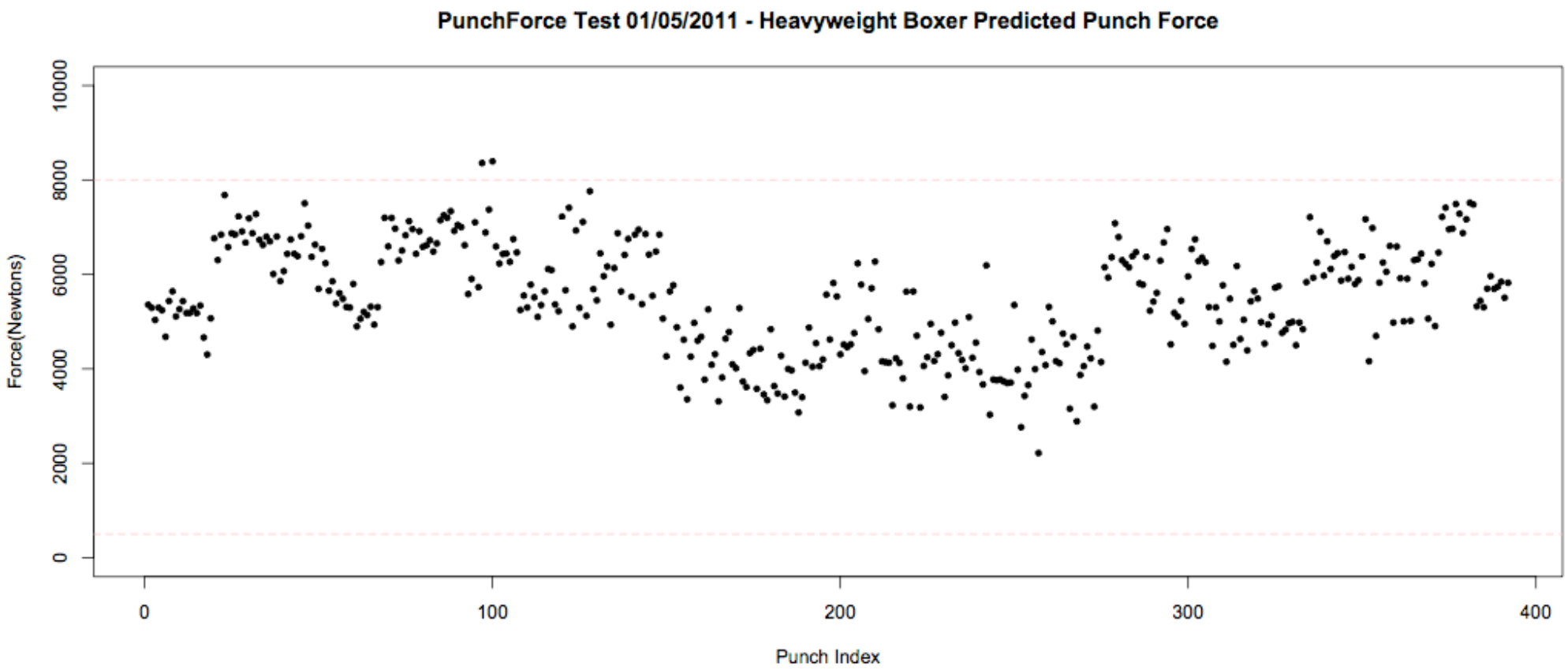


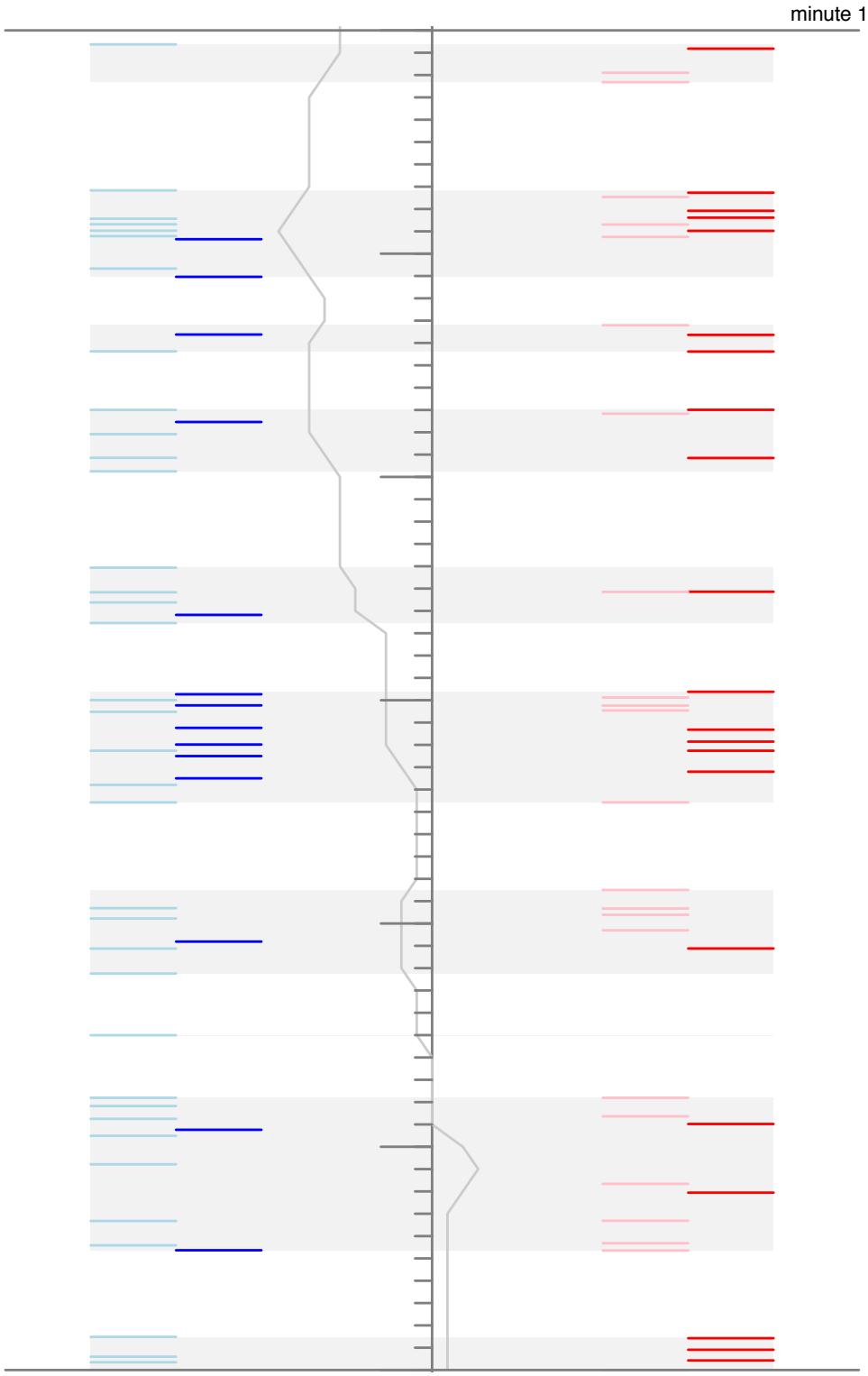

Right
Acceleration(G's)
30
20
10
0
-10
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Time(MS)

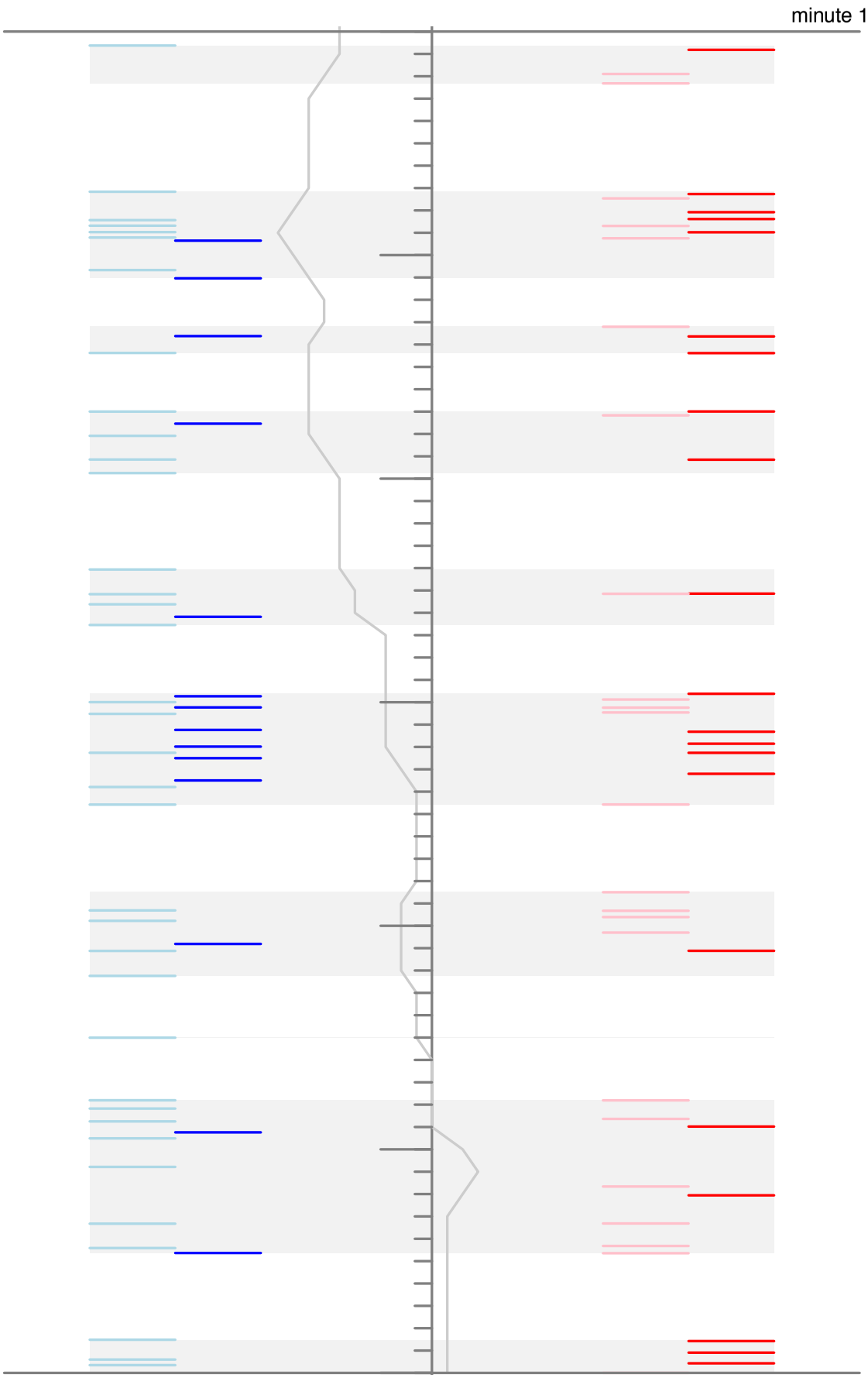

Right
Acceleration(G's)
30
20
10
0
-10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05

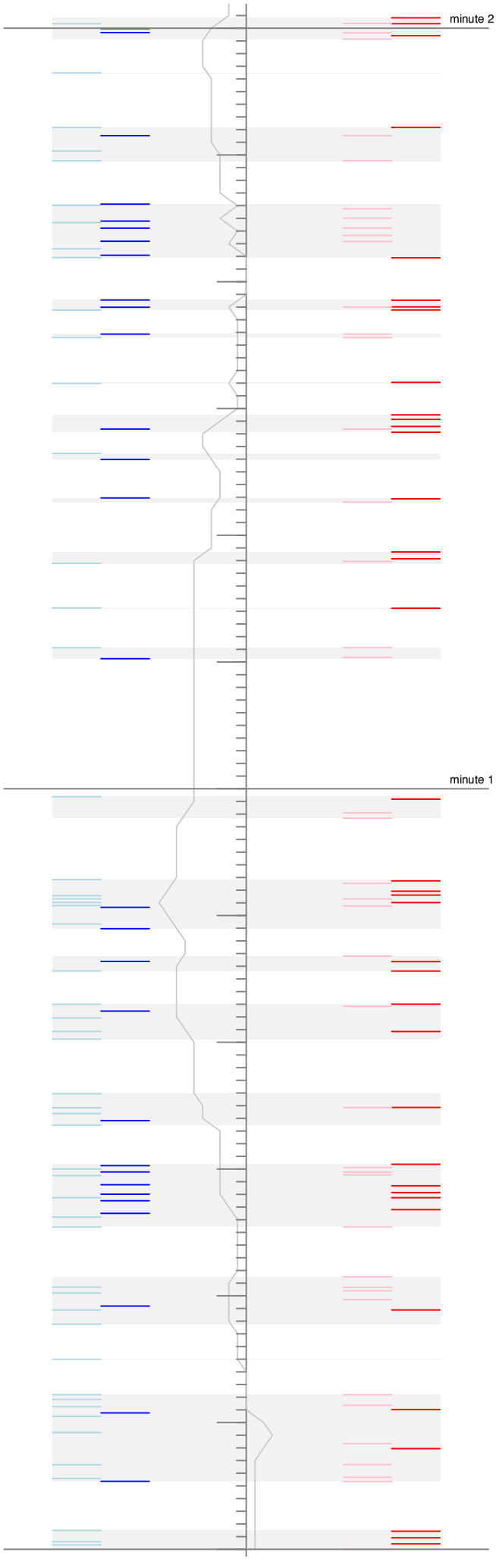

Velocity Calculation Using Acceleration Integral
Punch Impact
Punch Start
Acceleration (g's)
30
25
20
15
10
5
0
-5
-400
-200
0
200
400
Time(ms) from Peak Acceleration

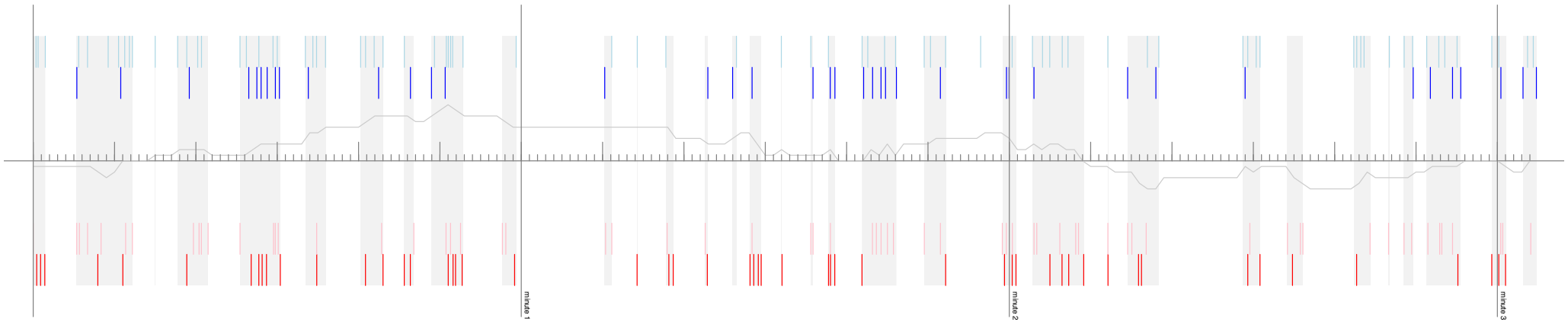
```
<?xml version="1.0" encoding="UTF-8"?>
<splineModel>
   <term>
      <coefficient>5817.4512456</coefficient>
      <se>2734.527438</se>
      <expression>5817.4512456</expression>
   </term>
   <term>
      <coefficient>365.2001095</coefficient>
      <se>32.998199</se>
      <product>
         <variable>
            <name>forearm.hit</name>
         </variable>
      </product>
      <expression>365.2001095*x[0]</expression>
   </term>
   <term>
      <coefficient>-423.9208580</coefficient>
      <se>77.889330</se>
      <product>
         <variable>
            <name>commitment</name>
         </variable>
      </product>
      <expression>-423.9208580*x[1]</expression>
   </term>
   <term>
      <coefficient>-682.2593177</coefficient>
      <se>58.719188</se>
      <product>
         <variable>
            <name>armweight.hit</name>
         </variable>
      </product>
```

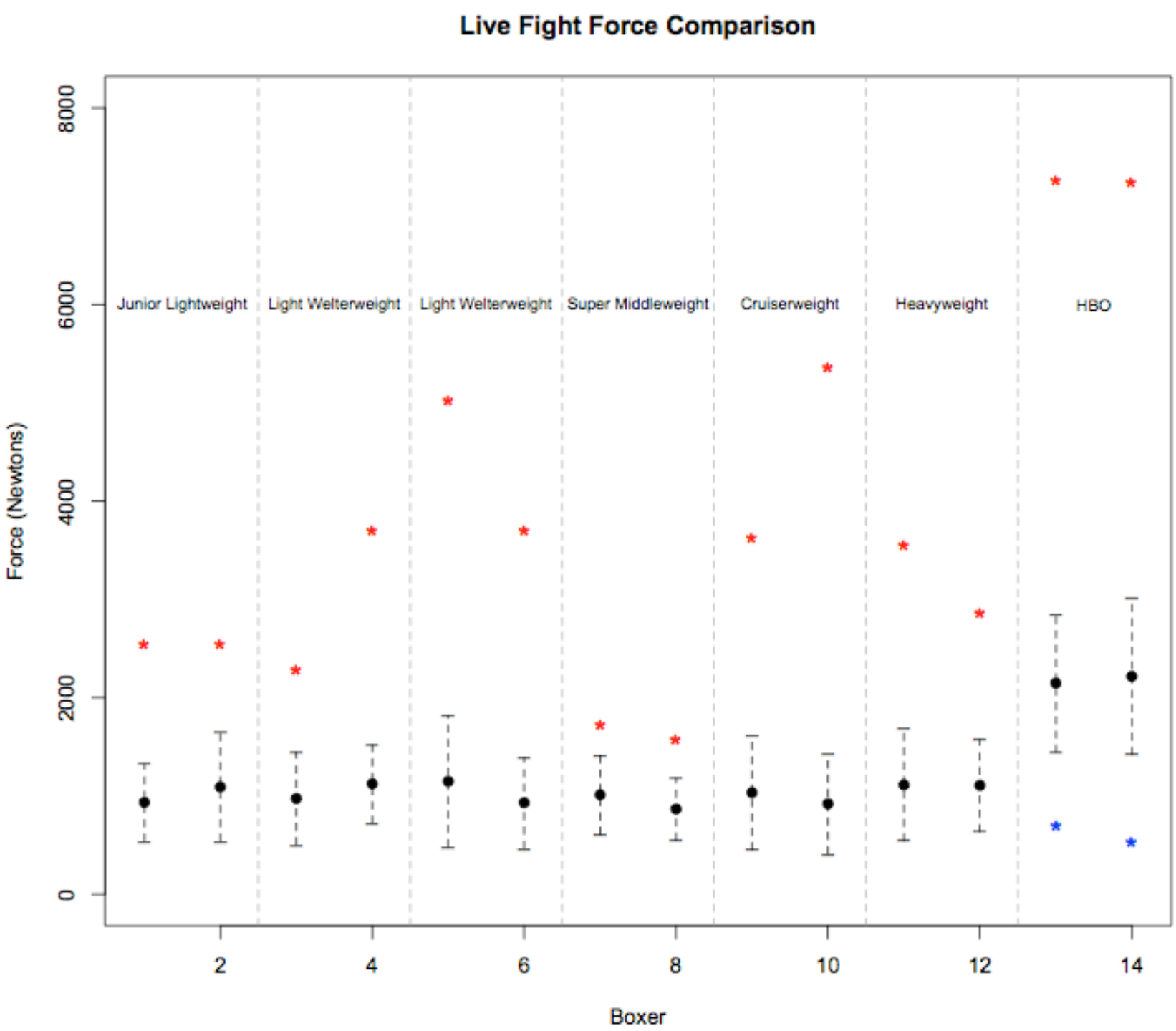

PunchForce Test 01/05/2011 - Heavyweight Boxer Predicted Punch Force
Force(Newtons)
0
2000
4000
6000
8000
10000
0
100
200
300
400
Punch Index


minute 1


minute 1


minute 2
minute 1


minute 1
minute 2
minute 3


Live Fight Force Comparison
Force (Newtons)
0
2000
4000
6000
8000
Junior Lightweight
Light Welterweight
Light Welterweight
Super Middleweight
Cruiserweight
Heavyweight
HBO
2
4
6
8
10
12
14
Boxer

Next time

We are going to start looking at **“statistical learning”** as a general topic -- We will see some simple modeling tools that let us understand the structure in a data set and even make predictions

We will revisit our ideas about testing and estimation throughout this discussion, and, in the end, provide you with **a broad understanding of the issues involved with learning patterns from data**

Inference

Here is a brief recap of where we have been so far...

1. We began the quarter by simply describing data, **describing the properties of samples**; we learned how to make particular kinds of plots and numerical summaries in R -- descriptors (formally, statistics) that helped reveal something of the patterns in our data, helping us **tell a story about the sample**

2. We then used many of these same descriptors (statistics) in a more formal way, creating **a simple framework for asking specific questions about how our data were generated**, for evaluating the plausibility of certain specific hypotheses -- we compared patterns evident in our data with **what should be expected under a particular "null" scenario**

3. As we worked through this second phase of the course, we started using descriptors (statistics) as more than just summaries; **we began to link them with populations, with the probability mechanism that gave rise to the data** -- slowly, we moved from description to estimation, and in so doing developed an interest in **the variability of our estimates**

Simulation

Throughout the quarter, computing generally and simulation in particular have played important roles

1. Through various examples, we have seen that a data set and its constituent parts are open to recoding, to transformation, to reshaping, to (re)combination; and that there is no natural "look" of a data set, merely appropriate choices of descriptors (numerical, graphical, statistics) that provide us a view of the data

2. In testing, **simulation was used to create distributions of our descriptors (statistics) that we would expect to see if a particular hypothesis were true;** in some cases we generated data from a particular model (say, tossing coins in the case of Arbuthnot's data), or we made use of special symmetries under the hypothesis being tested (say, a treatment has no effect, and hence we can ignore the given randomization and generate new ones)

3. In estimation, **simulation helped us assess the uncertainty in an estimate;** through the bootstrap, we estimated the sampling distribution and created confidence intervals for population parameters -- but there is a lot more to say on this point...

Estimation

Over the last two lectures, we have been developing tools for the following **simple estimation problem**:

We are interested in some **feature of a large population** and, for the purposes of this course, the feature needs to be represented numerically; that is, **we can assign some value to each item or subject in the population**

We then define some **parameter over the population values** referred to (originally enough) as θ; examples might include the average BMI of adult males in the US, "greenness" of limes in a particular chain of supermarkets, or the proportion of visitors who will respond to a particular web page layout

Finally, **we collect a simple random sample of items or subjects** from the population and examine their particular values; we let $\widehat{\theta}$ refer to **an estimate of the population parameter based on this sample**

Estimation

Over the last two lectures, we have been developing tools for the following **simple estimation problem**:

We then considered the sampling distribution of $\widehat{\theta}$; **it captures the variation we would expect to see if we repeated our experiment many times** (remember, we are in a frequentist mode where probability emerges from repeated trials)

For a given known population, the sampling distribution describes how far from θ the estimates $\widehat{\theta}$ are likely to be if we repeat our experiment; if we know how far $\widehat{\theta}$ tends to be from θ , then we know how far θ tends to be from $\widehat{\theta}$ **and confidence intervals are born**

This is a clean construction, but **it depends on us knowing the population**, but if we knew that, then there would be no need for taking a sample in the first place; **the bootstrap provided us an easy way approximate the sampling distribution in a large number of cases**

Simulation

The guiding mantra for this class might be "**analyze as you randomized**"; whether we were testing hypotheses about the effectiveness of some treatment and **re-randomizing** the results to creating confidence intervals by **re-sampling** through the bootstrap, the underlying probability mechanism that generated the data was front and center

An unexpected byproduct of our computational or simulation-based approach, then, is that **we are put face-to-face with the randomization** (whether random assignment to treatment and control or random sampling from a population) followed when the data were created

And that's a good thing...

Another approach

Of course, the bootstrap is a fairly new innovation, dating to the late 1970s -- performing this kind of analysis **before the days of a reasonably fast computer would be unthinkable**

Prior to this, however, there was (and continues to be) a thriving statistics industry that works with other techniques for approximating the sampling distribution **besides the bootstrap**

Let's consider one of the first...

The Central Limit Effect

We have made passing references to **the Central Limit Theorem** several times during this class; it is the magic in the background responsible for many of the bell-shaped distributions we've seen (either as reference distributions when we re-randomized or as estimates of the sampling distribution when we resampled)

Consider a simple estimate, the mean; to put this in the framework we just discussed, we are interested in **some average quantity taken over the entire population and will estimate it with the sample mean**

Some precise language

So far, we have been very loose about our language, using the terms “normal” and “bell-shaped” interchangeably; from now on, we will refer to normal distributions by name and not as “bell-shaped”

No matter what term we have been using, the normal shape said very specific things about where we could find data; the distribution is symmetric, centered on the mean and there are various “rules of thumb” like about 95% of the data are within 2 standard deviations of the mean, 99% within 3 and so on

At the end of the lecture, we will encounter another distribution, Student’s t-distribution, that looks a bit like a bell, but is not normal; it arranges data in a different way (it has heavier tails) so that you will see less that 95% within two standard deviations of the mean, for example

So, to avoid confusion, we will now use the term normal when we mean the normal distribution

The Central Limit Effect

Let μ be the population mean and suppose we collect n data points $x_1, x_2, \ldots, x_n$ from a simple random sample of the population, and let

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

be our sample-based estimate of μ -- Then, the Central Limit Theorem says that the sampling distribution of $\bar{x}$

1. Has mean μ ; that is, it's **centered on the true population parameter**

2. Its **standard error (the standard deviation of the sampling distribution) is** $\sigma/\sqrt{n}$, where σ is the standard deviation computed for the population

3. Has roughly **a normal distribution**, providing the sample size, n, is not too small

The Central Limit Effect

Remember we alluded to the fact that the term “normal” in this class will apply in the sense of **“conforming to a rule or pattern”**; here this pattern involves the shape of the sampling distribution as your sample size increases

To give you a sense of how this works, **we again appeal to simulation**; but as with our other sampling distribution simulations, we have to pretend we know the truth -- this is a magical state of being that we slip into to illustrate this effect

Simulation

It is another use for simulation: To verify properties of estimates under known conditions, and in this case, **to test out a mathematical result about an estimate that is beyond the scope of this class to formally prove**

To be clear, however, **we will frame the slides in cyan when our simulations are based on knowing the truth** -- this will hopefully reduce confusion when we return to estimation and resampling

With those caveats in place, how should we proceed? What needs to be verified? How will we go about checking?

Simulation

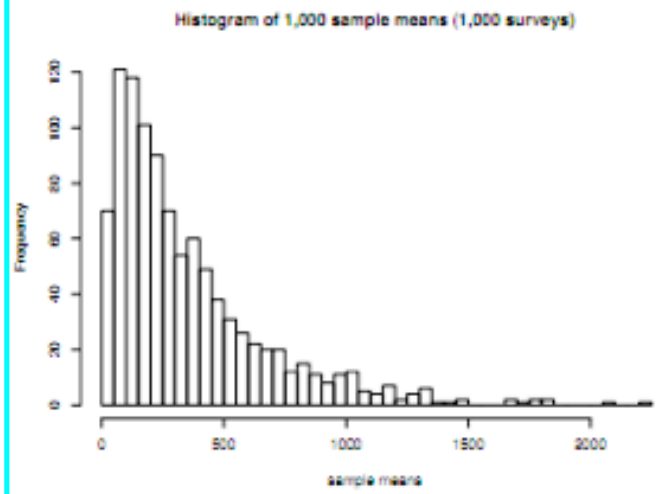
On the next three slides, we sample from the most twisted population we can imagine, the amount of time visitors spend on the NY Times travel section -- remember that had an intensely long right tail

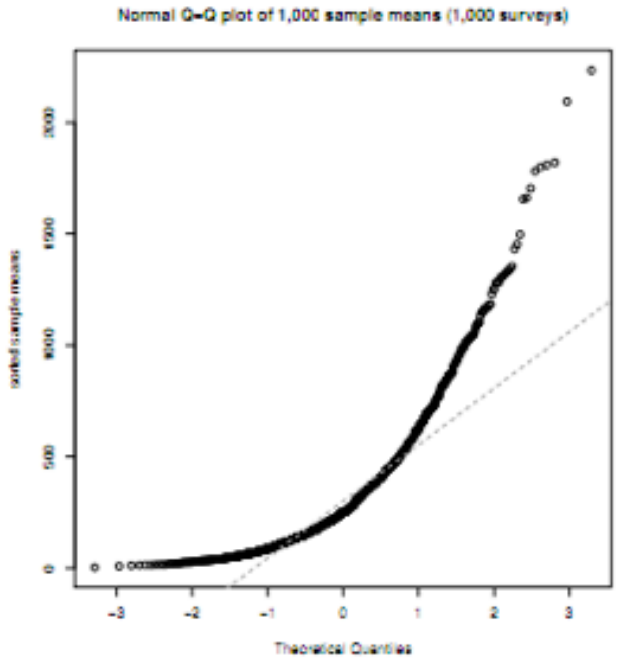
To be very clear, our POPULATION will now be the collection of 47,000 NY Times visitors we looked at a few lectures ago; we will take SAMPLES from this population and study the sampling distribution of $\bar{x}$

Our sample sizes increase from 2 to 5 to 10 to 50 to 100 to 250 to 500 and finally to 1000; with each we present two kinds of plots

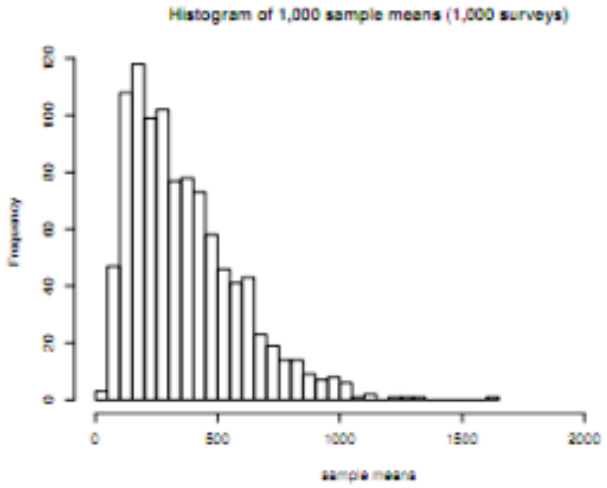
What do you see?

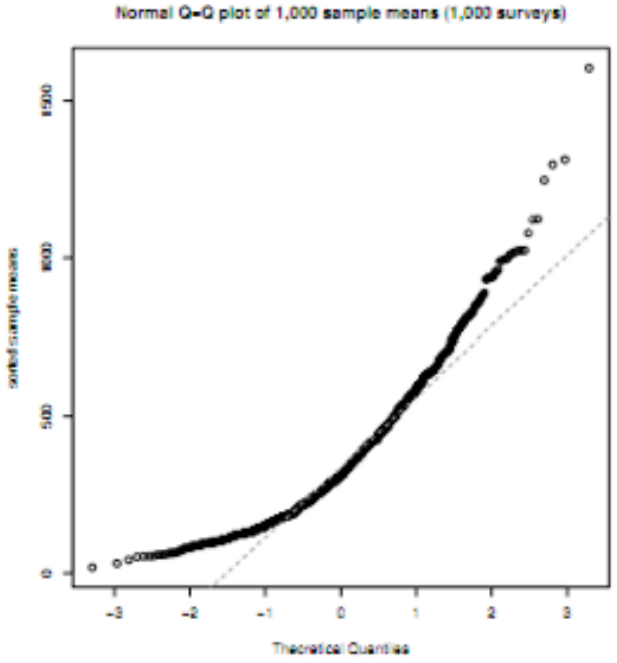
Surveys of size n=2



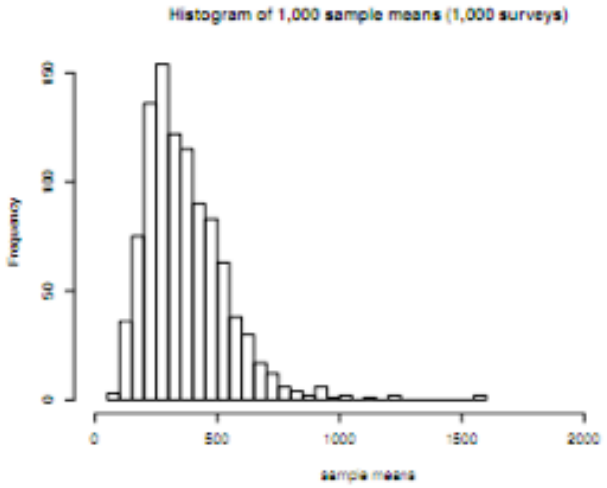


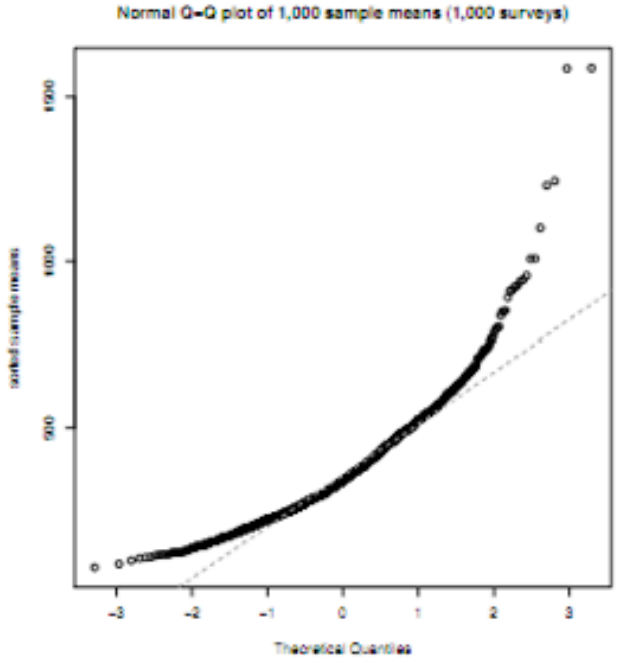
Surveys of size n=5



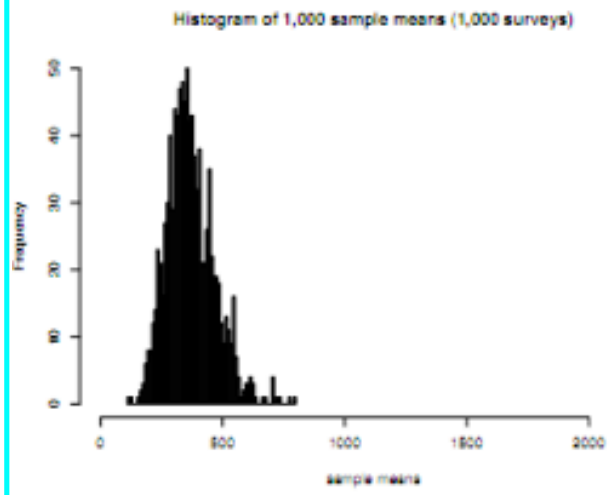


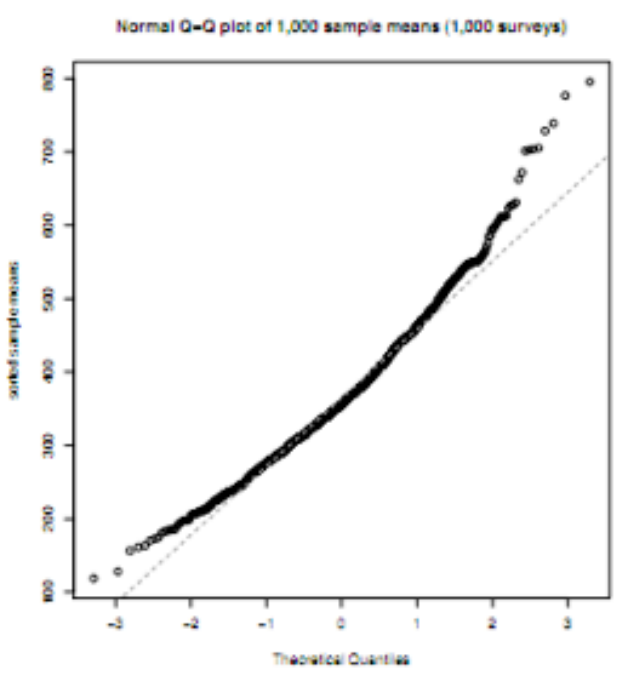
Surveys of size n=10



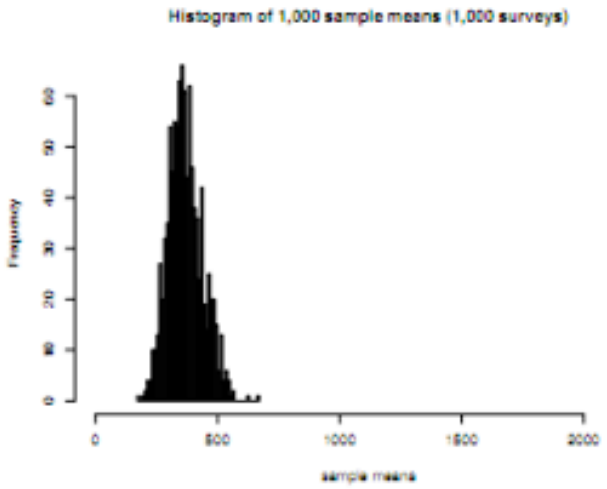


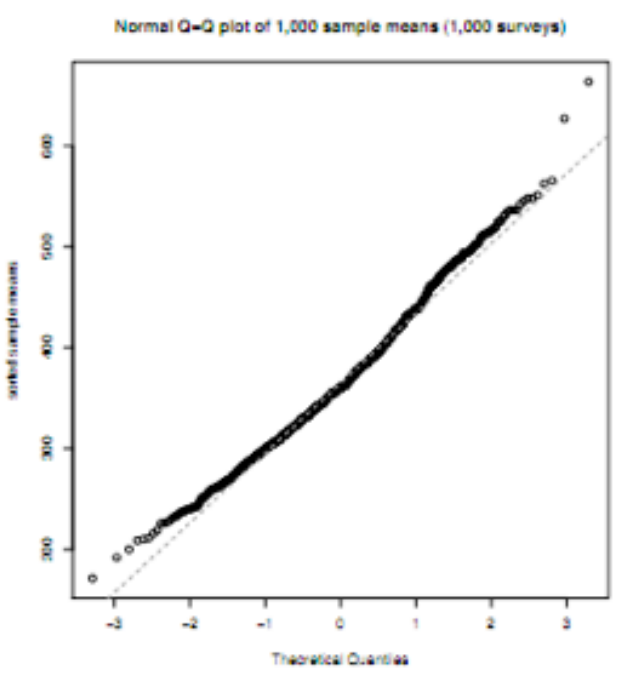
Surveys of size n=25



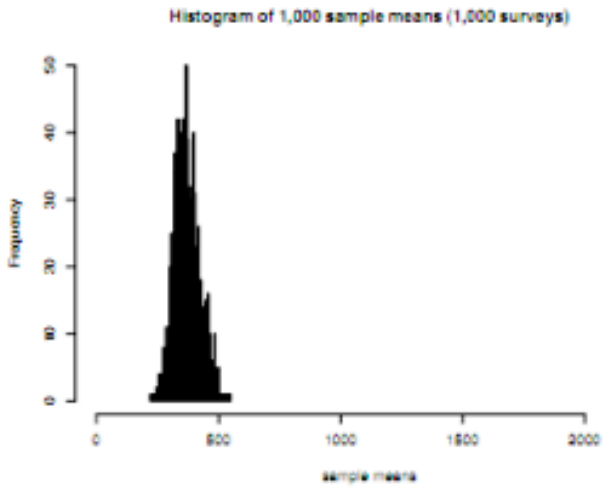


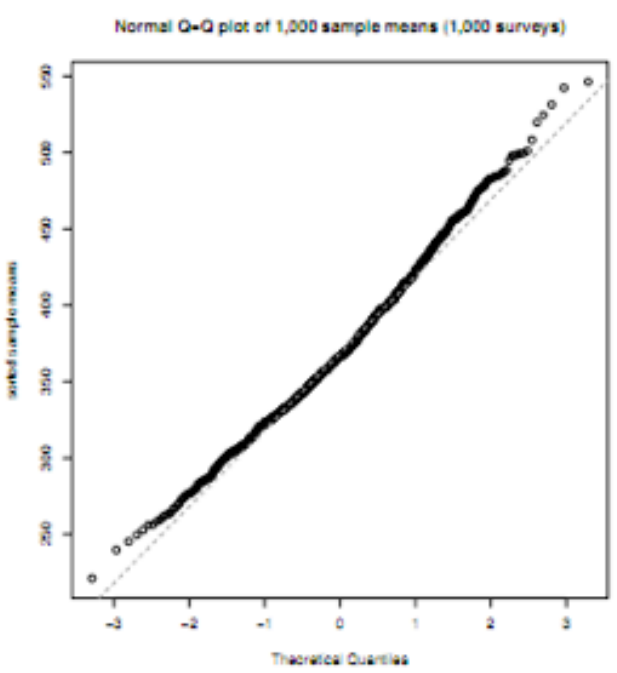
Surveys of size n=50



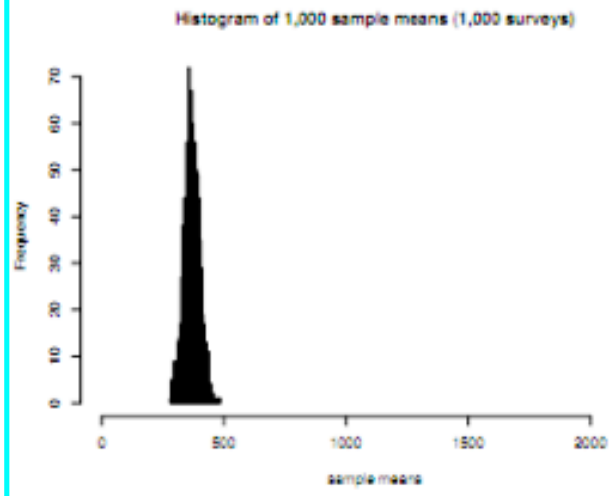


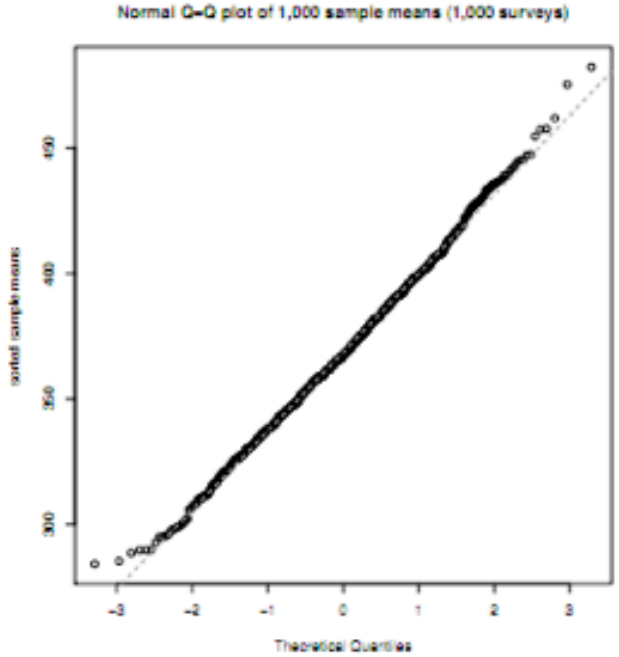
Surveys of size n=100



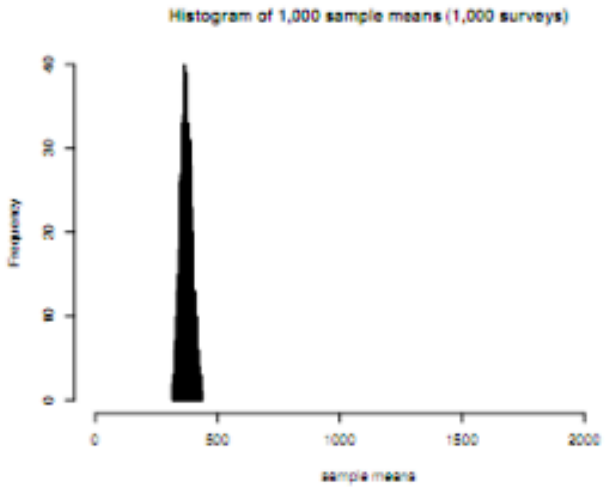


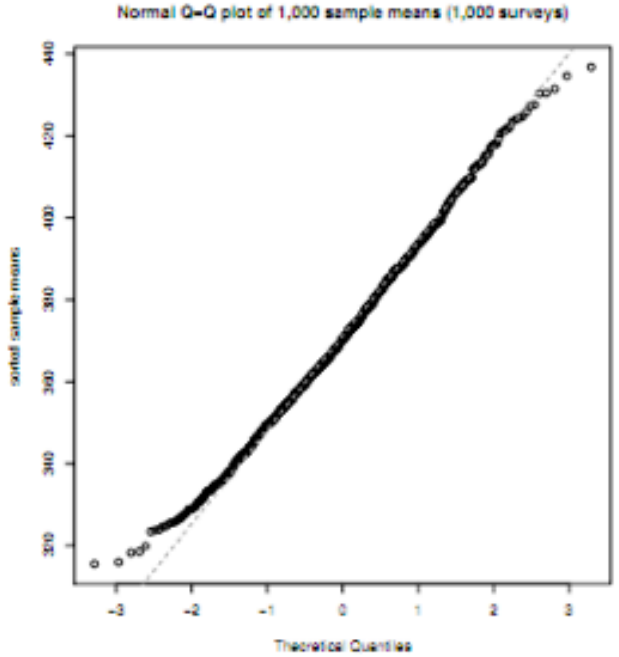
Surveys of size n=250



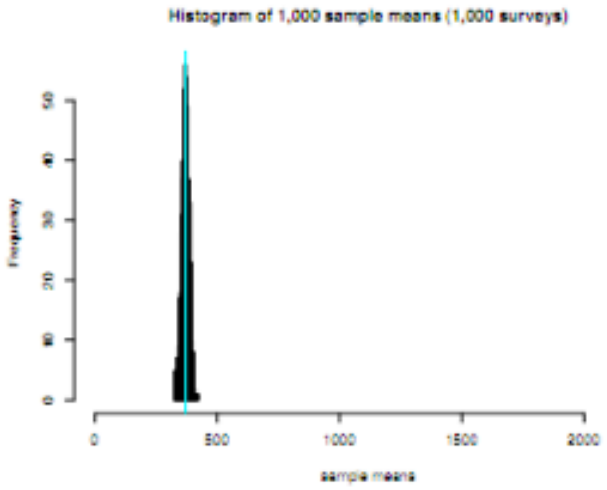


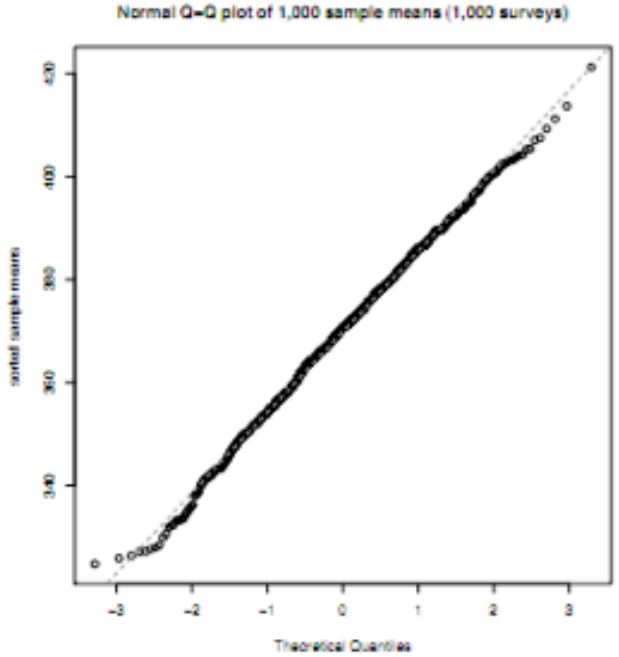
Surveys of size n=500





Surveys of size n=1,000





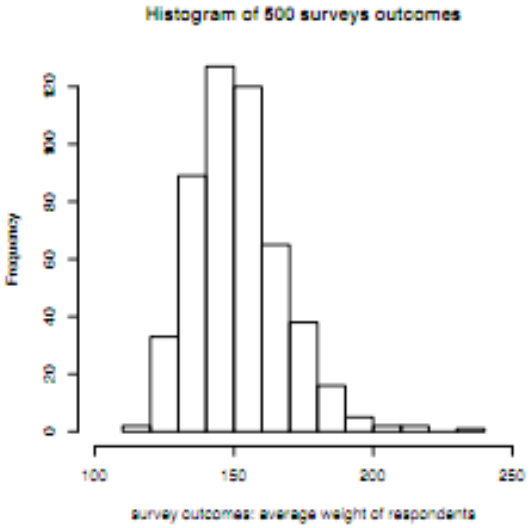
Simulation

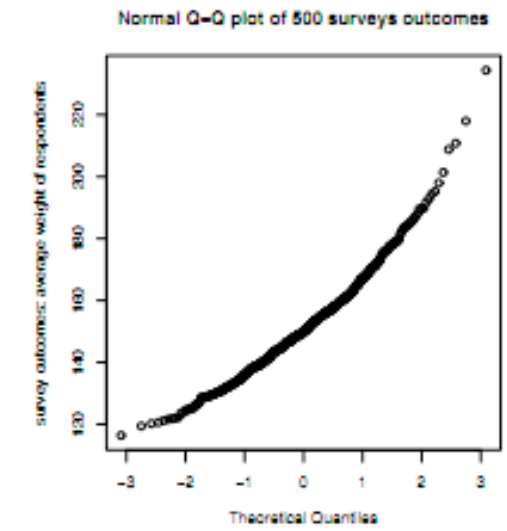
Now, let's try something a bit tamer; instead of the horrible distribution of visit times, let's instead consider using the CDC data set as a POPULATION, and focus on people's weights

Recall that the weight of people in the study had some skew (a heavy right tail -- although not nearly as bad as the NY Times data); we will take SAMPLES from this POPULATION and consider the sampling distribution of $\bar{x}$

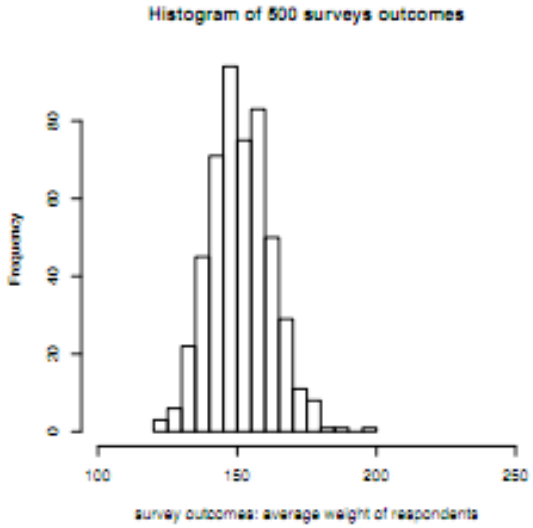
This time we consider samples of size n=5, 10 and 25 -- What do you see? Is there a difference in behavior from the last example? Why?

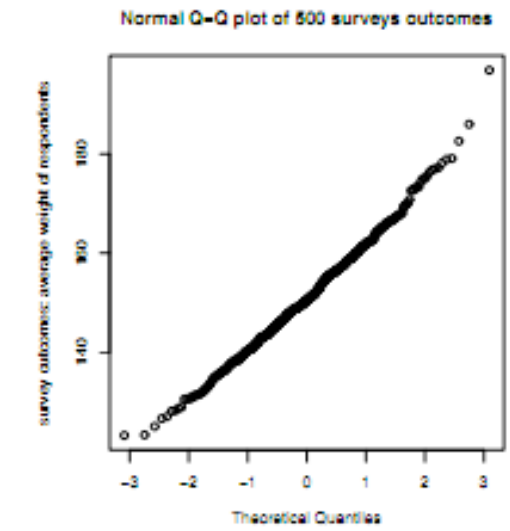
Average weight of respondents for surveys with n=5



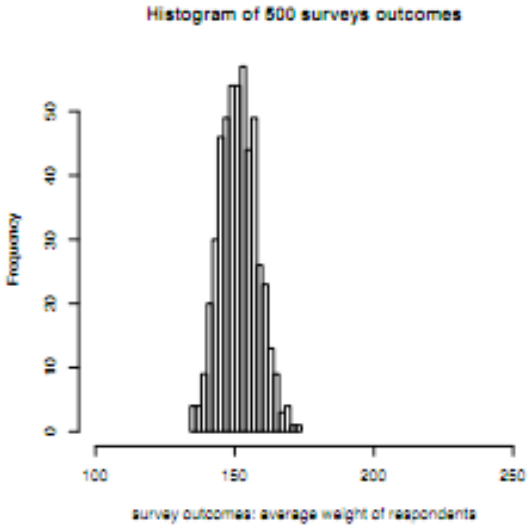


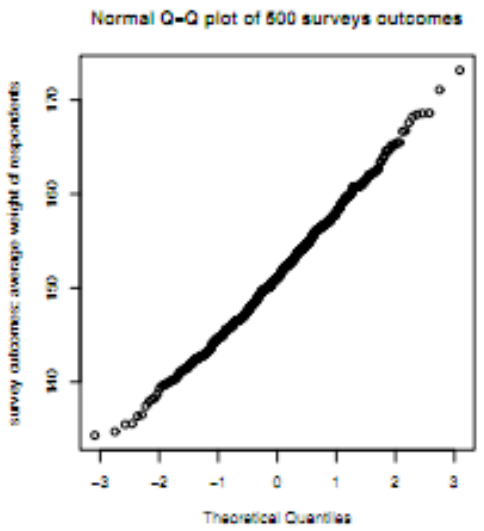
Average weight of respondents for surveys with n=10





Average weight of respondents for surveys with n=25





A traditional confidence interval

The behavior of the sample mean has been used by statisticians for quite some time; the result says that **if our sample size is "large enough",** then we can expect a normal sampling distribution that is centered on the population mean and its standard error is $\sigma/\sqrt{n}$

Put another way, we can form a rough 95% confidence interval starting with

$$\text{Prob}\left(\mu - 2\sigma/\sqrt{n} < \bar{x} < \mu + 2\sigma/\sqrt{n}\right) \approx 0.95$$

or, which by pivoting (if the estimates are close to the population mean, then flipped around, the mean must be close to the estimates) yields

$$\text{Prob}\left(\bar{x} - 2\sigma/\sqrt{n} < \mu < \bar{x} + 2\sigma/\sqrt{n}\right) \approx 0.95$$

This all looks great; to form a 95% confidence interval we add and subtract twice the standard error (the population standard deviation divided by the square root of n) **-- Wait, is there a problem here?**

A traditional confidence interval

The obvious problem is that we don't know σ ; or, I should say, that in almost every case where we are taking samples from a population, if we don't know the mean μ we won't know σ

So, now what?

The plug-in principle

When you don't know some population quantity, **the plug-in principle says you, well, plug in an estimate from the sample**

The term "plug-in principle" was coined (or so I have been told) by the creator of the bootstrap, Brad Efron (a statistician at Stanford University); and in some sense you can think of **the bootstrap as the ultimate plug-in -- we don't know the population so we are going to plug-in the sample!**

Following this, then, you have the traditional formula for a 95% confidence interval (for large sample sizes) $\bar{x} \pm 2s/\sqrt{n}$ where s is the sample standard deviation defined a few lectures ago

A comparison

Before you ask, it turns out that the bootstrap estimate of the standard error (the one you we computed by repeatedly sampling from the distribution) is essentially the same as $s/\sqrt{n}$ -- that is, $sd(\widehat{\theta}^*)$ (the standard deviation of our bootstrap replicates) we talked about in lecture and you will compute in lab is essentially $s/\sqrt{n}$

That means, **for large sample sizes** where we have normal sampling distributions, **every scheme you've seen so far for computing a confidence interval should agree** (that's comforting, right?)

Um, so why the bootstrap?

Rationale

For many (many many) statistics we are interested in, we don't have a formula for the standard error like we do for the sample mean; **with the bootstrap, no formula is needed** -- and in the few cases such a thing exists, it will agree with the classical formula

Instead of dealing in formulae, we rely on R or some other bootstrap enlightened software package to provide us with **a ready assessment of the precision of our estimate** -- it is fully general and quite powerful

Of course standard errors are only part of the game; we can also examine the bias of an estimate and we can generate confidence intervals that are free of the normal assumption (recall our percentile approach)

Back to the sample mean

So far, we have been leaning quite heavily on the Central Limit Theorem, and there are two open issues: The first is that we don't know how big a sample size has to be for a normal distribution to appear, and the second is that we don't know σ (we plugged in s, but does that "work"?)

For the bootstrap this isn't a problem because we would appeal to the percentile approach (or it's souped up cousin) if we noticed something strange going on; in Lab this week you will work with the bootstrap package in R and see how all this is implemented

In the late 1800s one statistician began to investigate one of these problems in earnest; his approach would have a huge impact on the field


GUINNESS®
ST. JAMES'S GATE BREWERY, DUBLIN

Some history

Guinness Brewing Company incorporated in 1886 and was soon the largest brewery in the world, delivering 1.5 million barrels a year in England, Ireland and around the world

At the time, brewers learned “meticulously the traditional practices” as apprentices; the Chairman and Managing Director at Guinness wanted to change all this **-- they wanted to make brewing scientific... and they invested heavily in the idea**

They started by **hiring top-notch chemists from Oxford and Cambridge** Universities; at a rate of one every one or two years starting in 1893 -- these chemists began projects to, for example, “identify and quantify what it was that gave hops and barley their brewing qualities”

This group **examined all aspects of production**, from the “raw materials” to how best to cultivate, fertilize, dry and store the barley

Some history

Guinness supported this group as they ran agricultural experiments (first renting, then buying farms) to acquiring an experimental malthouse to finally a separate brewery they could use to conduct their experiments

"Reading led to analysis, experiments and measurements. They began to accumulate data and, at once, ran into difficulties because their measurements varied. The effects they were looking for were not usually clear cut or consistent, as they had expected, and they had **no way of judging whether the differences they found were effects of treatment or accident**"

In October of 1899 Wiliam S. Gosset, a recent chemistry graduate from Oxford was hired into this group; he had studied some mathematics and so this group brought him their data to analyze -- Gosset wrote "It may seem strange that reasoning of this nature had not been widely made use of, but this is due, first, to the popular dread of mathematics"

Some history

Gosset realized that in the cases he was facing (small samples, high variation), the usual Central Limit Theorem didn't apply, and he couldn't ignore the fact that σ was being estimated

Gosset decided to study the sampling distribution for the sample mean, or rather a "standardized" quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

focusing, in particular, on cases for small values of n, but **assuming the population itself was normal**

His approach was novel; he decided to come up with **an exact expression for the sampling distribution, but under a strict assumption about the population** (one that he felt matched the experimental conditions he was seeing)

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

By Student

Gosset's research was conducted while he was employed at Guinness; the Managing Director agreed that he could publish his results but added "that such publication might be made without the brewer's name appearing -- instead Guinness employees would be designated **'Pupil' or 'Student'"**

While we mentioned that Gosset studied a little mathematics, **he was by no means a mathematician**; his paper includes a couple insightful moves that have little to do with formal proof

So, it's 1905, say, and you want to study a sampling distribution and the mathematics is beyond you: What do you do? Hint: What would you do now?

LITERATURE.

(1) Frew, J. G. H. (1923). On the larval anatomy of *Chlorops taeniopus* Meig. and two related Acalyptrate Muscids, with notes on their winter host plants. (*P.Z.S.* p. 783.)

(2) Fulmek, L. (1911). Zum Auftreten der Halmfliege (*Chlorops taeniopus* Meig.) in Weizen. (*Osterreichischen Agrar-Zeitung* Nr. 30 vom 29 Juli.)

(3) Nowicki, M. (1871). Ueber die Weizenverwusterin *Chlorops taeniopus* Meig. und die Mittel zu ihrer Bekampfung. (*Verh. Zoologischbotanischen Gesellschaft in Wien.*)

(4) Ormerod, E. A. (1890). Manual of Injurious Insects and Methods of Prevention (2nd ed. London, pp. 75–79).

(5) "Mathetes" (1924). Statistical study on the effect of manuring on infestation of barley by Gout Fly. (*Ann. App. Biol.* xi, 2.)

By Student

You simulate! Reading over the **first part of Section VI**, you can see how painful this task was; it involved creating a physical version of the population, with values written on cards

Ah, but another question arises: It is 1905 and you need observations from a normal distribution; Where do you turn?

Now 50 to 1 corresponds to three times the probable error in the normal curve and for most purposes would be considered significant; for this reason I have only tabled my curves for values of n not greater than 10, but have given the $n=9$ and $n=10$ tables to one further place of decimals. They can be used as foundations for finding values for larger samples*.

The table for $n=2$ can be readily constructed by looking out $\theta=\tan^{-1}z$ in Chambers' Tables and then $\cdot5+\theta/\pi$ gives the corresponding value.

Similarly $\frac{1}{2}\sin\theta+\cdot5$ gives the values when $n=3$.

There are two points of interest in the $n=2$ curve. Here s is equal to half the distance between the two observations. $\tan^{-1}\frac{s}{s}=\frac{\pi}{4}$ so that between $+s$ and $-s$ lies $2\times\frac{\pi}{4}\times\frac{1}{\pi}$ or half the probability, i.e. if two observations have been made and we have no other information, it is an even chance that the mean of the (normal) population will lie between them. On the other hand the second moment coefficient is

$$\frac{1}{\pi}\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}\tan^2\theta\,d\theta=\frac{1}{\pi}\Big[\tan\theta-\theta\Big]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}=\infty\,,$$

or the standard deviation is infinite while the probable error is finite.

Section VI. *Practical Test of the foregoing Equations.*

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell (*Biometrika*, Vol. I. p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation† of each sample determined. The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the z of Section III.

This provides us with two sets of 750 standard deviations and two sets of 750 z's on which to test the theoretical results arrived at. The height and left middle finger correlation table was chosen because the distribution of both was approximately normal and the correlation was fairly high. Both frequency curves, however, deviate slightly from normality, the constants being for height $\beta_1=\cdot0026$, $\beta_2=3\cdot175$, and for left middle finger lengths $\beta_1=\cdot0030$, $\beta_2=3\cdot140$, and in consequence there is a tendency for a certain number of larger standard deviations to occur than if the distributions were normal. This, however, appears to make very little difference to the distribution of z.

* E.g. if $n=11$, to the corresponding value for $n=9$, we add $\frac{1}{5}\times\frac{2}{6}\times\frac{3}{7}\times\frac{4}{8}\times\frac{5}{9}\cos^8\theta\sin\theta$; if $n=13$ we add as well $\frac{1}{10}\times\frac{2}{7}\times\frac{3}{8}\times\frac{4}{9}\times\frac{5}{10}\cos^{10}\theta\sin\theta$ and so on.

† I hope to publish the results of the correlation work shortly.

By Student

You simulate! Reading over the **first part of Section VI**, you can see how painful this task was; it involved creating a physical version of the population, with values written on cards

Ah, but another question arises: It is 1905 and you need observations from a normal distribution; Where do you turn?

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing ***the height and left middle finger measurements of 3000 criminals****, from a paper by W. R. Macdonell. The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains t****he measurements of 3000 criminals in a random order****. Finally* ***each consecutive set of 4 was taken as a sample*** *- 750 in all - and the mean, standard deviation and correlation of each sample determined.* ***The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample...***

TABLE III. 3000 *Criminals.* Height (feet and inches).

Left Middle Finger (millimetres).	4′ 7″⅛—8″⅛	8″⅛—9″⅛	9″⅛—10″⅛	10″⅛—11″⅛	5′ 11″⅛—0″⅛	0″⅛—1″⅛	1″⅛—2″⅛	2″⅛—3″⅛	3″⅛—4″⅛	4″⅛—5″⅛	5″⅛—6″⅛	6″⅛—7″⅛	7″⅛—8″⅛	8″⅛—9″⅛	9″⅛—10″⅛	10″⅛—11″⅛	6′ 11″⅛—0″⅛	0″⅛—1″⅛	1″⅛—2″⅛	2″⅛—3″⅛	3″⅛—4″⅛	4″⅛—5″⅛	Totals
9·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
8	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
9	—	—	1	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
10·0	1	—	—	1	2	—	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	7
1	—	—	—	1	3	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	10
2	—	—	2	2	2	1	—	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	17
3	—	1	1	3	2	2	3	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	20
4	—	—	1	1	2	3	3	4	3	3	—	—	—	—	—	—	—	—	—	—	—	—	30
5	—	—	—	1	3	7	6	4	3	1	3	1	—	1	—	—	—	—	—	—	—	—	44
6	—	—	—	1	4	5	9	14	6	3	1	—	—	1	—	—	—	—	—	—	—	—	74
7	—	—	1	2	4	9	14	16	15	7	3	1	2	—	—	—	—	—	—	—	—	—	75
8	—	—	—	2	5	6	14	27	10	7	1	2	1	—	—	—	—	—	—	—	—	—	102
9	—	—	—	—	2	6	14	24	27	14	10	4	1	—	—	—	—	—	—	—	—	—	163
11·0	—	—	—	2	6	12	15	31	37	27	17	10	6	—	—	—	—	—	—	—	—	—	152
1	—	—	—	3	3	12	22	26	24	26	24	7	4	1	—	—	—	—	—	—	—	—	183
2	—	—	—	3	2	7	21	30	38	29	27	20	4	1	—	—	—	—	—	—	—	1	164
3	—	—	—	1	—	5	10	24	26	39	26	24	7	2	—	—	—	—	—	—	—	—	228
4	—	—	—	—	3	4	9	29	56	58	26	22	10	11	—	—	—	—	—	—	—	—	233
5	—	—	—	—	—	5	11	17	33	57	38	34	25	11	2	—	—	—	—	—	—	—	226
6	—	—	—	—	2	1	4	13	37	39	48	38	27	12	2	2	—	—	—	—	—	—	232
7	—	—	—	—	—	2	9	17	30	37	48	45	24	9	2	2	—	1	—	—	—	—	184
8	—	—	—	—	1	—	2	11	15	35	41	34	29	10	5	1	—	—	—	—	—	—	162
9	—	—	—	—	1	1	2	12	10	27	32	35	19	10	9	3	1	—	—	—	—	—	163
12·0	—	—	—	—	—	—	1	4	8	19	42	39	22	16	8	2	2	—	—	—	—	—	126
1	—	—	—	—	—	—	—	2	4	13	22	28	15	27	10	4	1	—	—	—	—	—	91
2	—	—	—	—	—	—	1	2	5	6	23	17	16	11	8	1	1	—	—	—	—	—	89
3	—	—	—	—	—	—	—	—	4	8	10	13	20	23	6	5	—	—	—	—	—	—	44
4	—	—	—	—	—	—	1	1	1	2	7	12	4	7	7	1	—	—	1	—	—	—	52
5	—	—	—	—	—	—	—	1	—	1	3	12	11	8	6	8	—	2	—	—	—	—	35
6	—	—	—	—	—	—	—	—	—	1	—	3	5	7	8	6	3	1	1	—	—	—	31
7	—	—	—	—	—	—	—	—	—	1	1	7	5	5	8	2	2	—	—	—	—	—	25
8	—	—	—	—	—	—	—	—	—	—	1	2	3	1	8	5	3	1	1	—	—	—	7
9	—	—	—	—	—	—	—	—	—	—	—	1	2	2	—	1	1	—	—	—	—	—	8
13·0	—	—	—	—	—	—	—	—	—	—	3	—	1	—	1	—	2	1	—	—	—	—	2
1	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	1	—	—	—	—	6
2	—	—	—	—	—	—	—	—	—	—	1	1	—	1	—	3	—	—	—	—	—	—	2
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	0
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	
Totals	1	1	6	23	48	90	175	317	393	462	458	413	264	177	97	46	17	7	4	0	0	1	3000
Means	100	103	102·8	107·0	107·8	109·4	110·6	111·8	113·3	114·8	116·5	117·7	118·6	120·1	122·2	123·9	125·9	126·4	127·7	—	—	112	

To clarify why measurements of criminals were in the "public data domain"
From the first page of the article...

PART I.

Material and Methods.

(1) The object of this memoir is threefold:

(i) To test to what extent the criminal classes diverge in physical characters from other classes of the community.

(ii) To consider how far the shorter methods recently proposed by Professor Karl Pearson for finding the variability and correlation of characters in the case of normal frequency may be applied to some of the chief anthropometric measurements now customarily made, and

(iii) To determine what is the best manner in which these measurements can be applied to the identification of criminals.

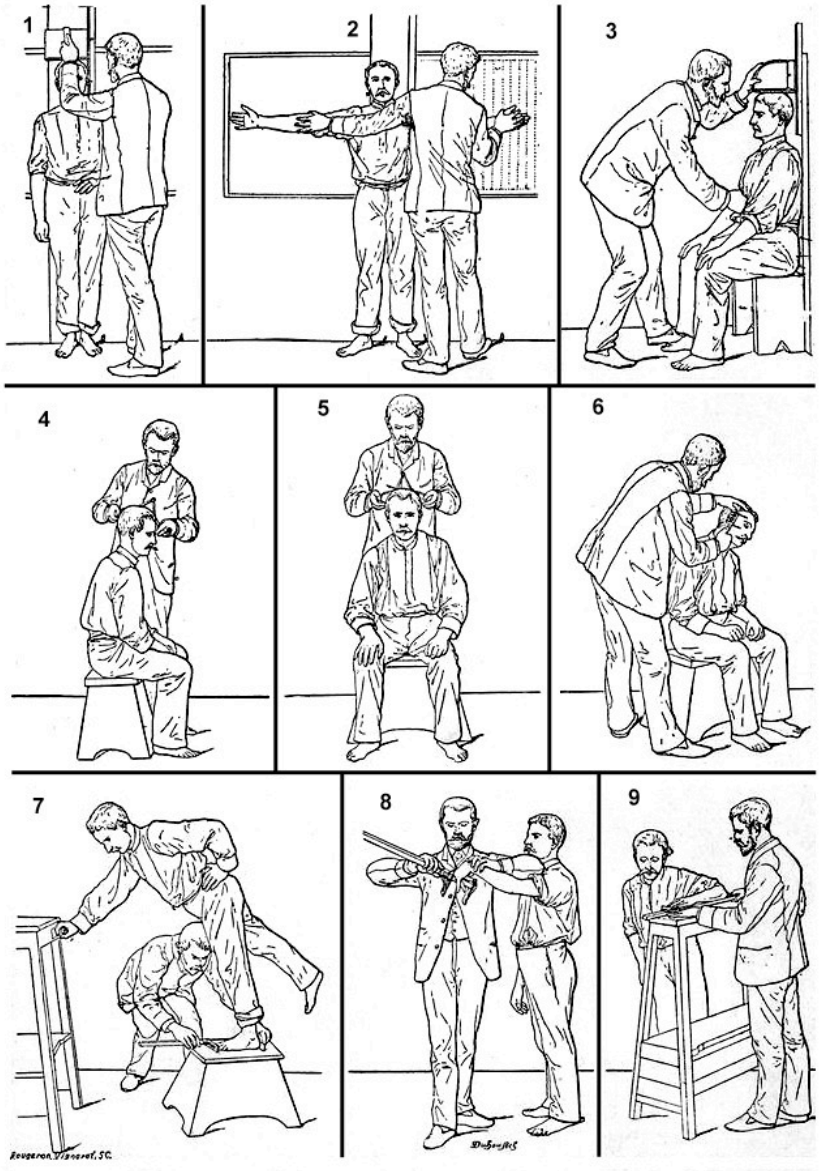
And later some conclusions...

Summing up the results of this part of the inquiry, I conclude that there is a substantial difference in stature, and in size and shape of head between the two classes; I do not assert that the source of the criminality is to be found in this difference, but only that criminals are drawn from a different section of the community. As bearing on this point it is worth noting that the mean height in Galton's middle-class measurements at the International Exhibition of 1884, viz. 67″·9, approaches our criminal mean more closely than does the Cambridge mean.

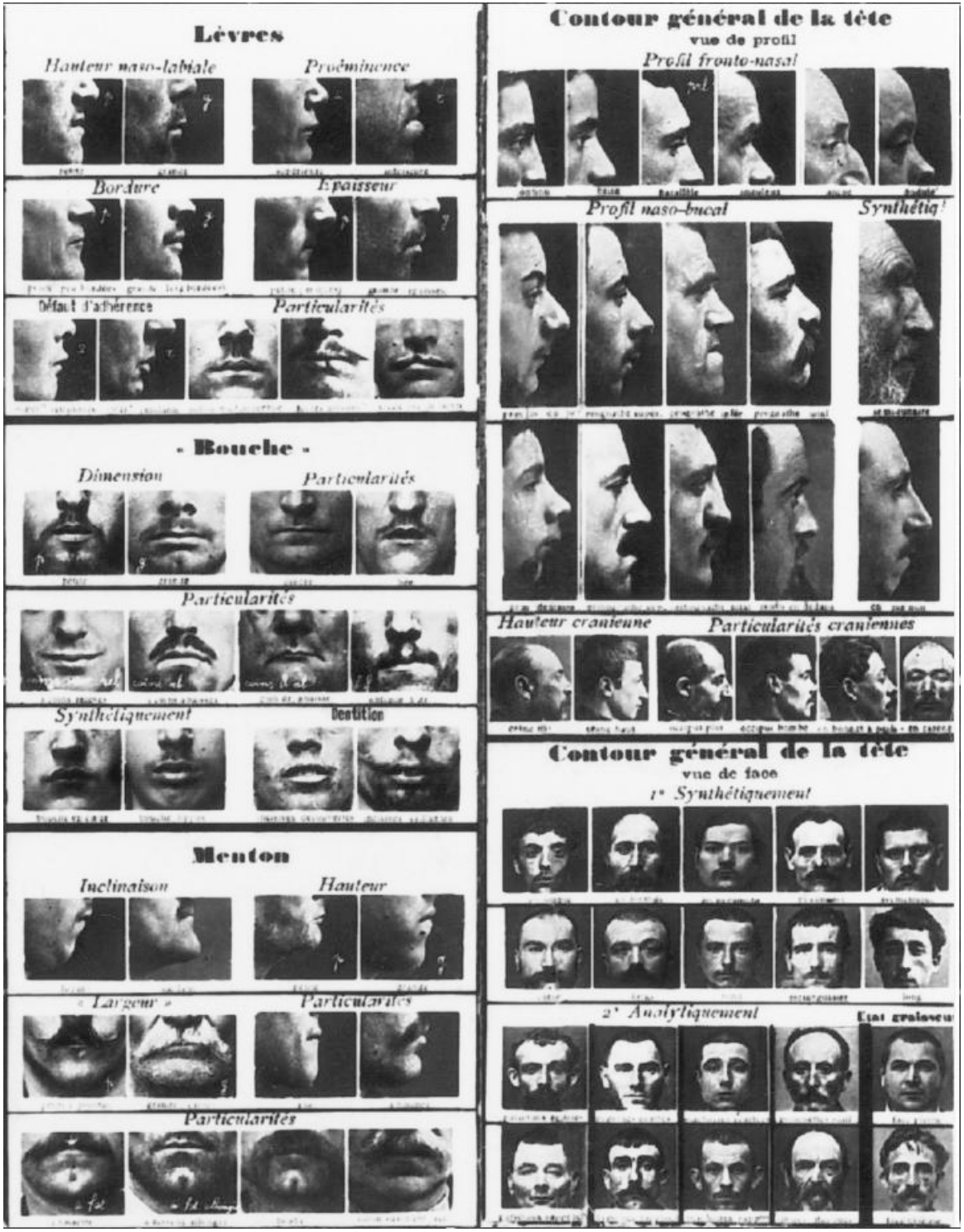
From "On Criminal anthropometry and the identification of criminals" by W. R. MacDonell, *Biometrika* 1, Vol 1, 1902

"Every measurement slowly reveals the workings of the criminal. Careful observation and patience will reveal the truth."

Alphonse Bertillon
French criminologist

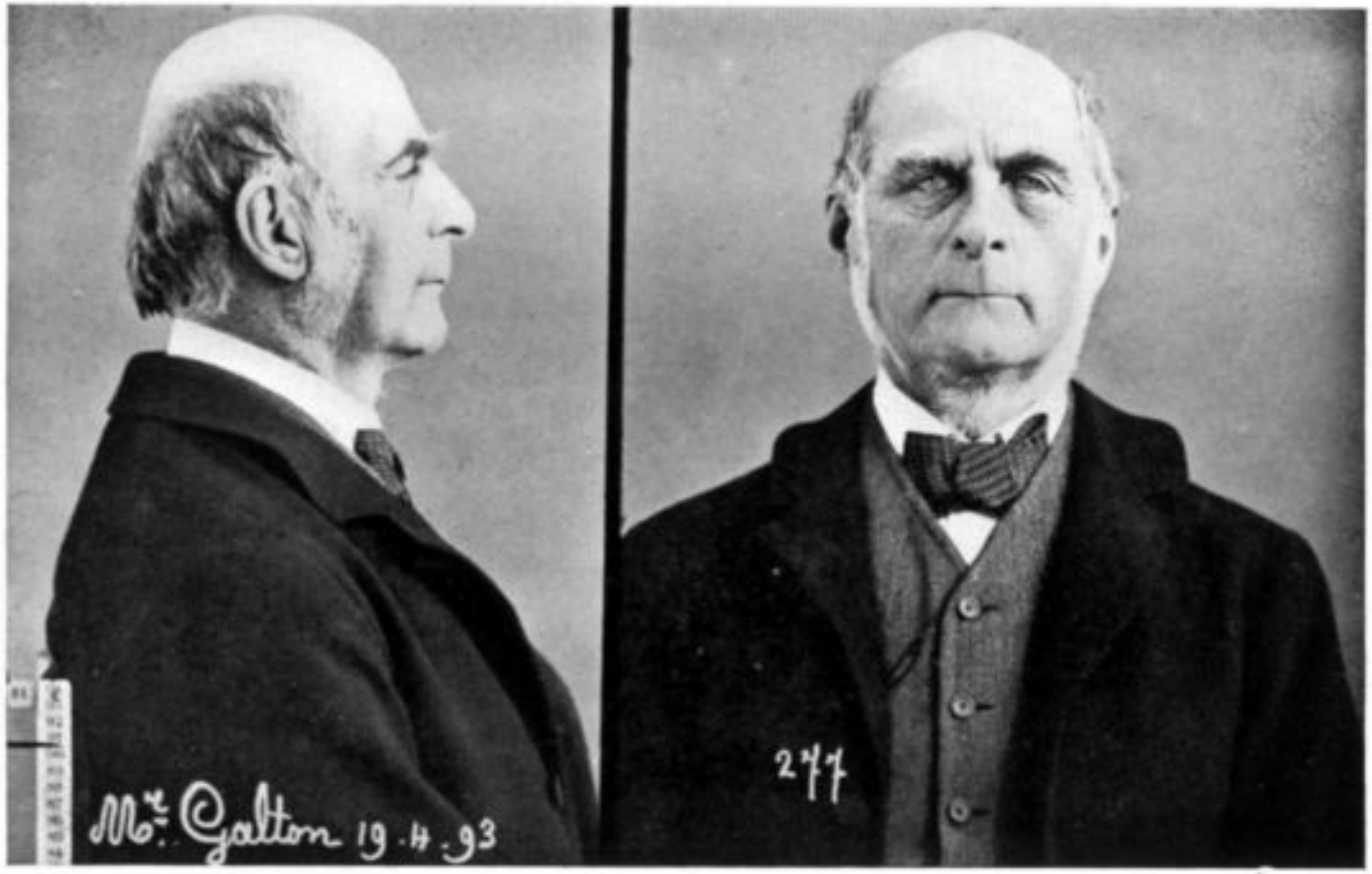


1. Height.
2. Reach.
3. Trunk
4. Length of head.
5. Width of head.
6. Right ear.
7. left foot.
8. Left middle finger.
9. Left forearm.


Lèvres
Hauteur naso-labiale
Proéminence
Bordure
Epaisseur
Défaut d'adhérence
Particularités
« Bouche »
Dimension
Particularités
Particularités
Synthétiquement
Dentition
Menton
Inclinaison
Hauteur
« Largeur »
Particularités
Particularités
Contour général de la tête
vue de profil
Profil fronto-nasal
Profil naso-bucal
Synthétiq!
Hauteur crânienne
Particularités crâniennes
Contour général de la tête
vue de face
1° Synthétiquement
2° Analytiquement
Etat graisseux

Taille 1m	Oreille dr. Tête. Long'	Pied g.	Coul' de l'iris g. N° de cl.	Agé de
Voute	Larg'	Médius g.	Aur'e	né le
Enverg 1m	Long'	Auric'e g.	Pér'e	à
Buste 0,	Larg'	Coudée g.	Part'e	dep't
				âge app't

(Réduction photographique 1/7.)

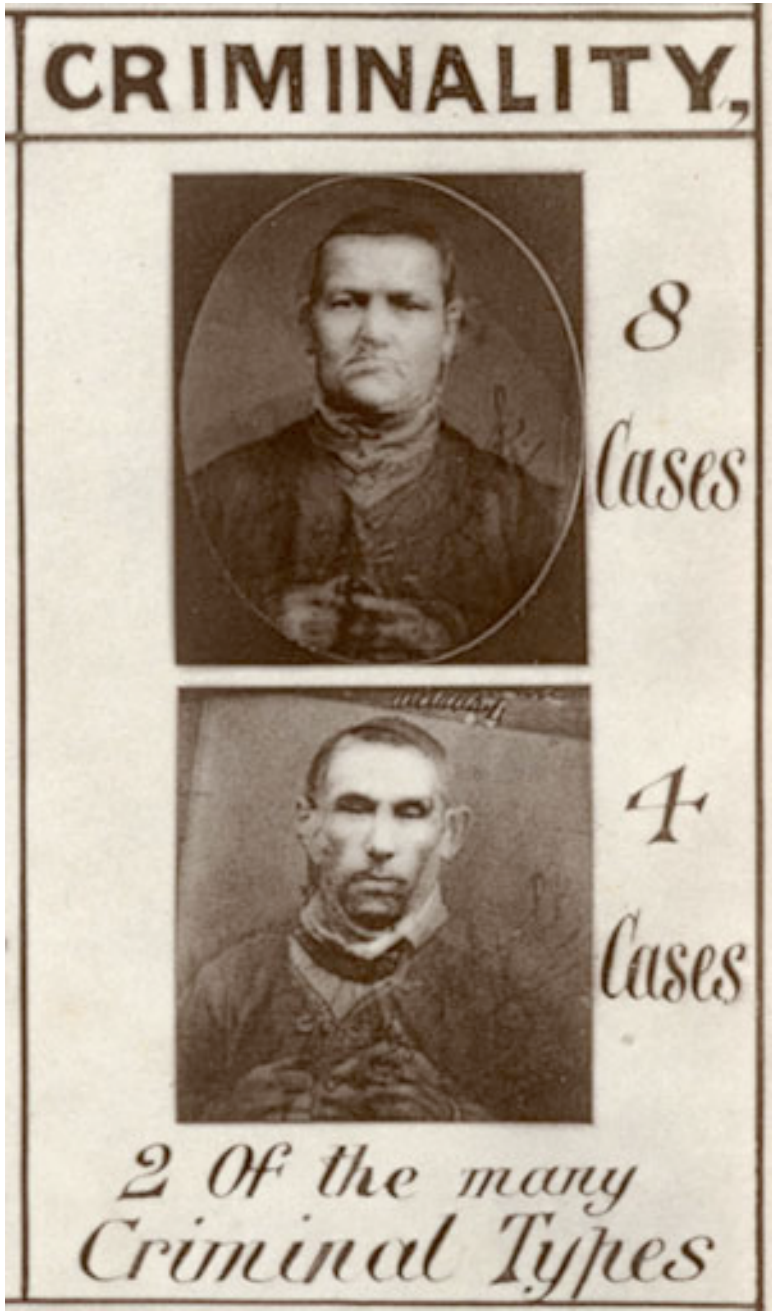


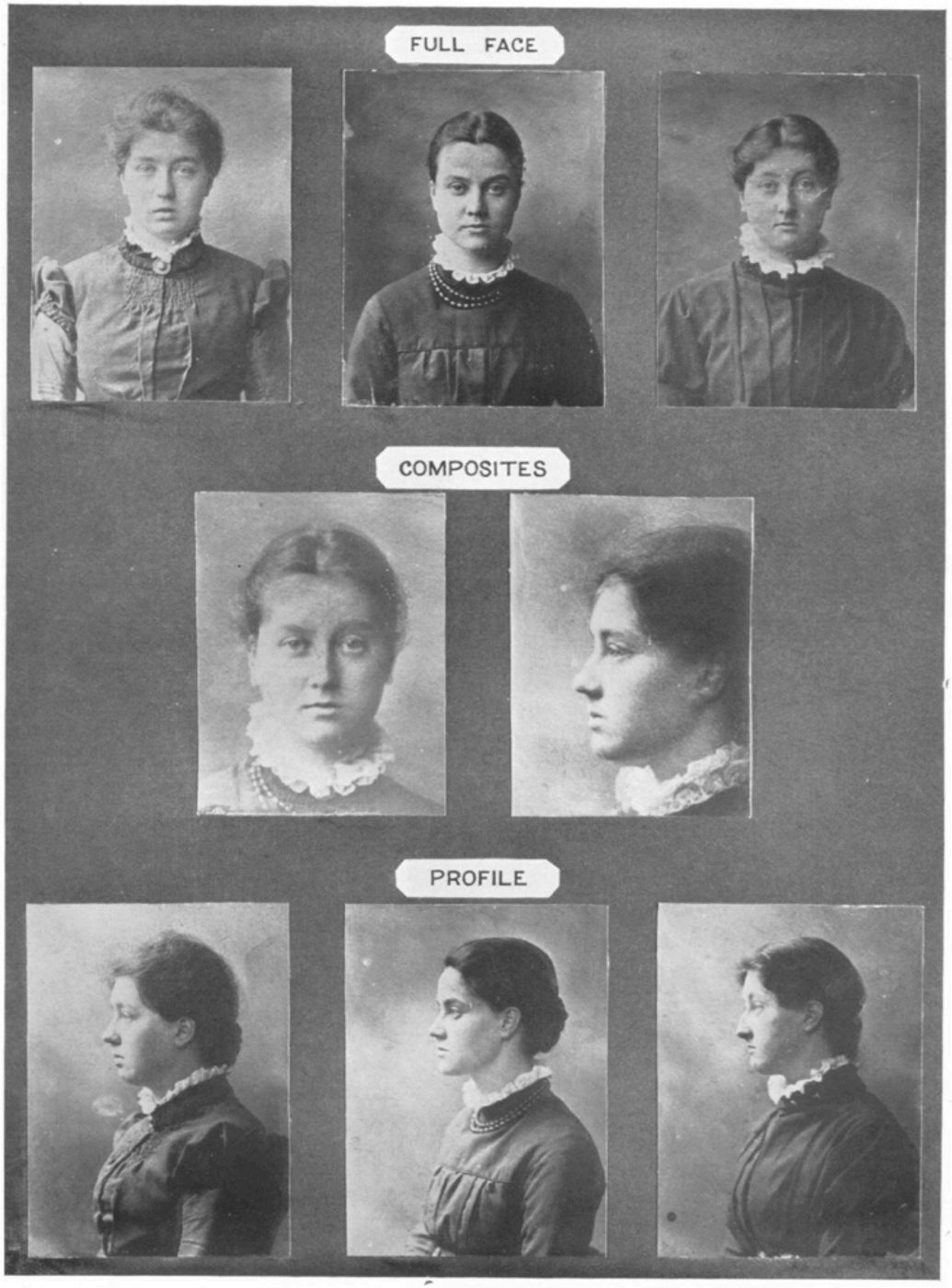
Galton

As we will see, Galton was deeply committed to the idea of **the normal curve as an important force in nature** and (as with Quetelet) thought the mean value had particular importance as **an indicator of "type"**

Quetelet was more extreme than Galton, however, in that he believed deviations from the mean were more like small errors, and **regarded the mean as something perfect or ideal**

For Galton, these types were stable from generation to generation -- You can see this in his work on fingerprints or even in his **composite photography**





Portraits of three Sisters, full face and profile, with the corresponding Composites.


5 COMPONENTS
7 COMPONENTS
4 COMPONENTS
2 COMPONENTS
PREVALENT TYPES OF FEATURES AMONG MEN CONVICTED OF LARCENY (WITHOUT VIOLENCE)

SPECIMENS OF COMPOSITE PORTRAITURE

PERSONAL AND FAMILY.

Alexander the Great From 6 Different Medals.

Two Sisters.

From 6 Members of same Family Male & Female.

HEALTH.

23 Cases. Royal Engineers. 12 Officers. 11 Privates

DISEASE.

6 Cases

9 Cases

Tubercular Disease

CRIMINALITY.

8 Cases

4 Cases

2 Of the many Criminal Types

CONSUMPTION AND OTHER MALADIES

I

20 Cases

II

36 Cases

56 Cases

Co-composite of I & II

Consumptive Cases.

100 Cases

50 Cases

Not Consumptive.

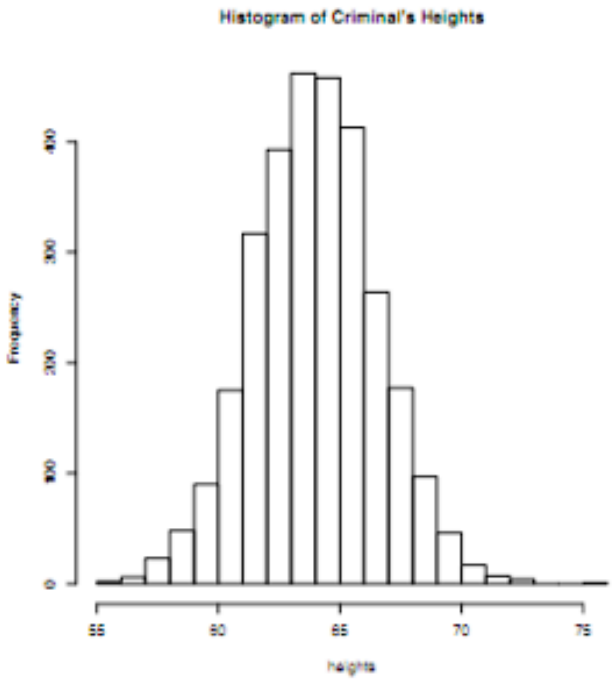
FIG. 9.—Enlarged impressions of the fore and middle finger tips of the right hand of Sir William Herschel, made in the year 1860.

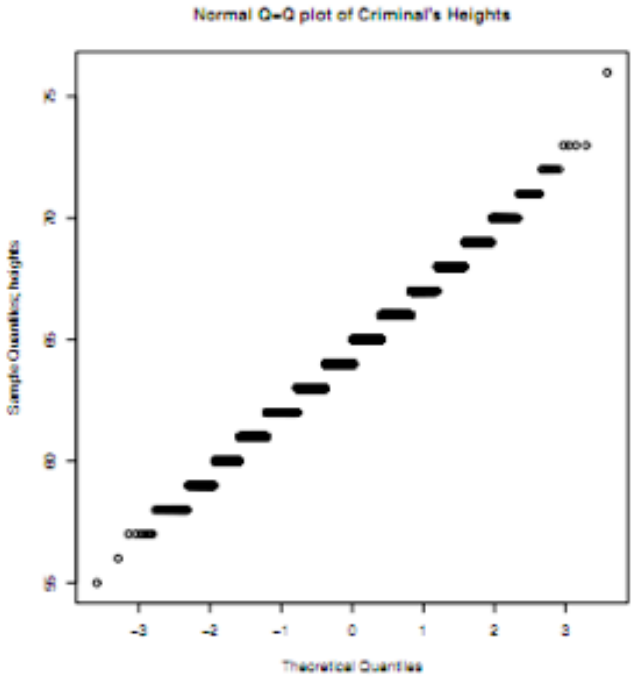
By Student

Here are Gosset's data using a couple of displays we're now very familiar with, a histogram and a normal Q-Q plot

Keep in mind these plots represent the ***entire population***; from this collection of 3,000 numbers we will draw samples (take surveys)

What do you notice?





By Student

We know the population $\mu = 64.5$ mean and the population standard deviation $\sigma = 2.6$

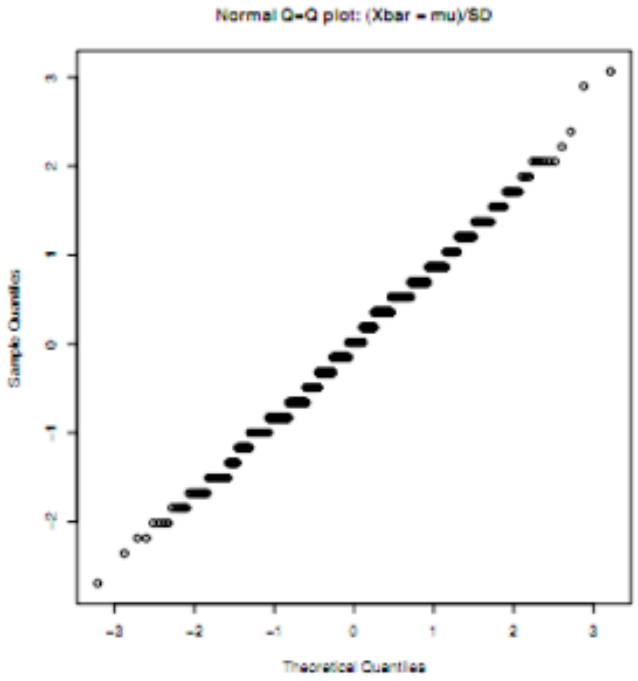
Gosset then took surveys of size *n*=4 and looked at

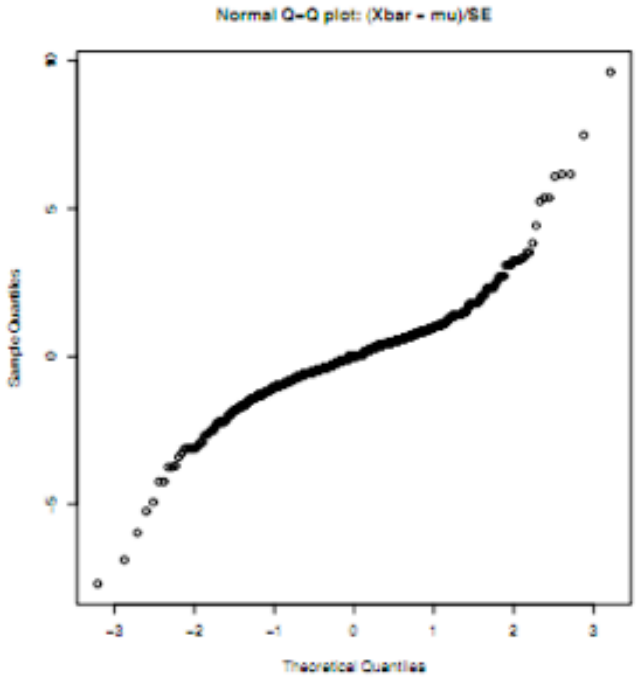
$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

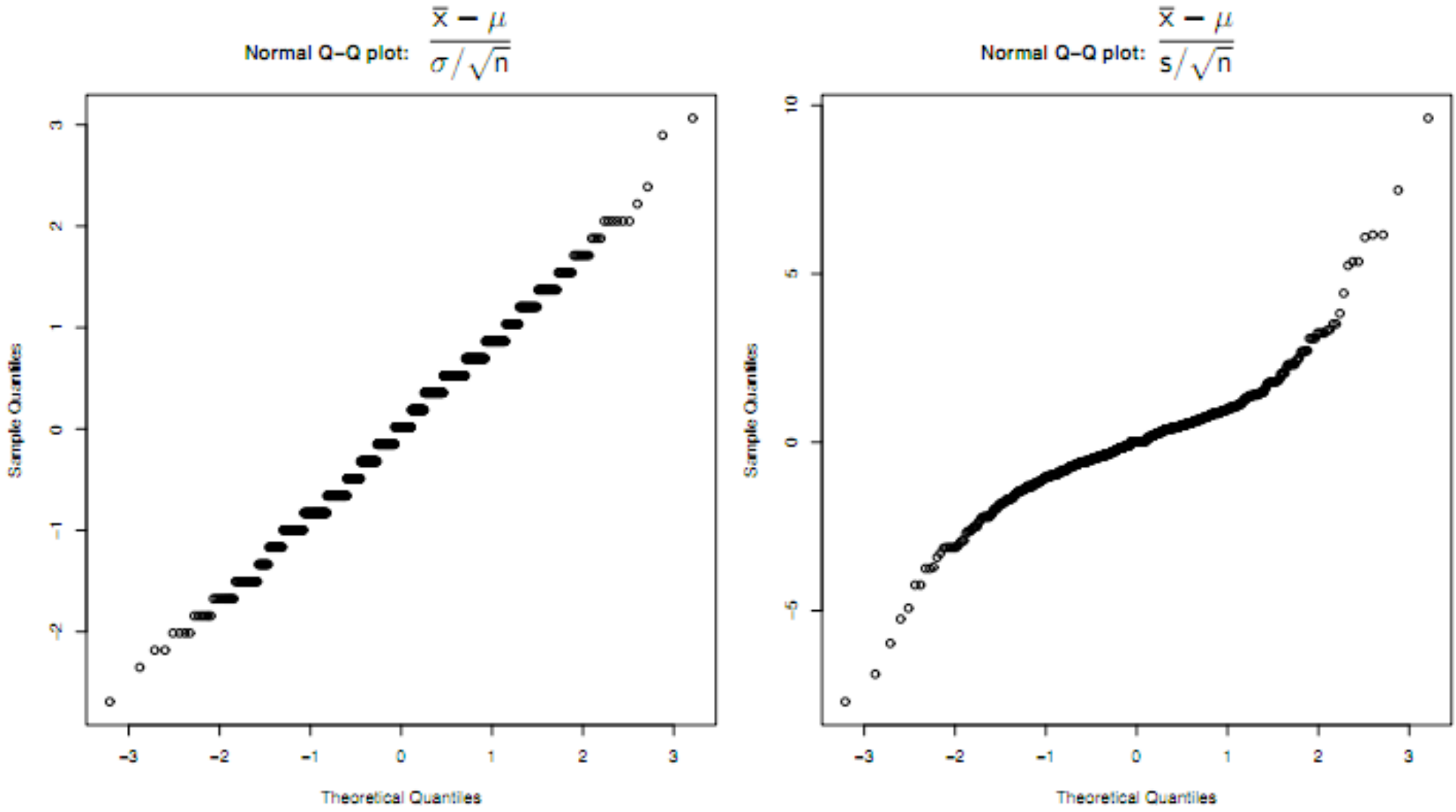
(top graph) and at

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

(bottom graph); What do you notice?







The effect of estimating σ, Gosset's simulation with sample size n=4: On the left he standardizes with the known population standard deviation and on the right he has "plugged-in" s for σ

The *t*-distribution

By having to estimate the population standard deviation in small samples, Gosset showed that the following equation have value somewhat less than 0.95

$$\text{Prob}\left(-2 < \frac{\bar{x}-\mu}{s/\sqrt{n}} < 2\right) = \text{Prob}\left(\bar{x} - 2s/\sqrt{n} < \mu < \bar{x} + 2s/\sqrt{n}\right)$$

The tails of the distribution of $(\bar{x}-\mu)/(s/\sqrt{n})$ are heavier than that of a normal; or, put another way, we see from the Q-Q plot both left and right skew

Intuitively, we have **a random quantity downstairs and this induces more spread in the distribution**

Gosset described the correct distribution when the feature of our population we're interested in has is normal looking to begin with (like heights); we refer to it as Student's *t*-distribution

The *t*-distribution

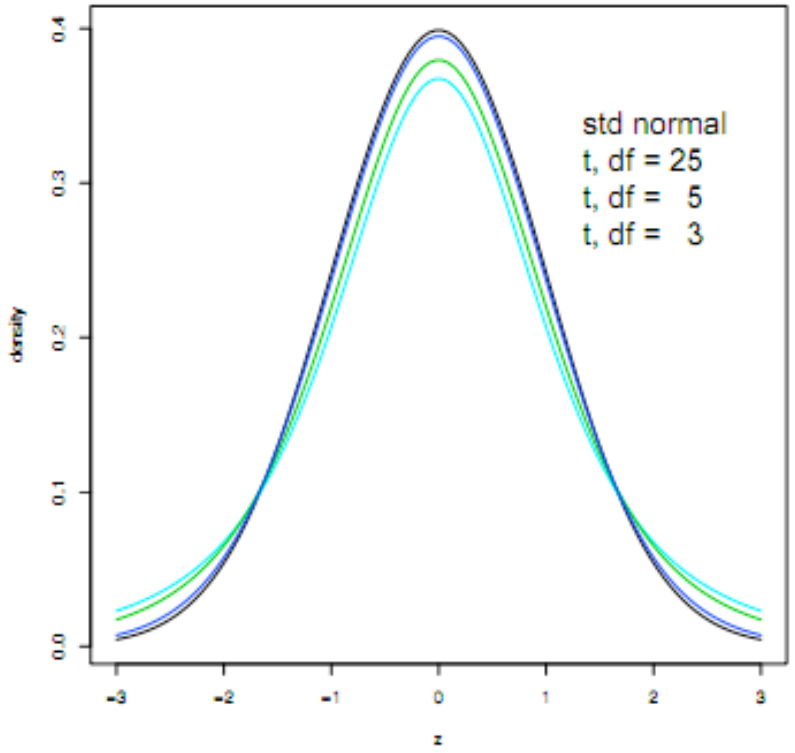
The *t*-distribution has one parameter controlling its shape; it is referred to as its *degrees of freedom*

In our context, the degrees of freedom is *n*-1, where *n* is the survey or sample size

It comes from our original definition of the sample standard deviation

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

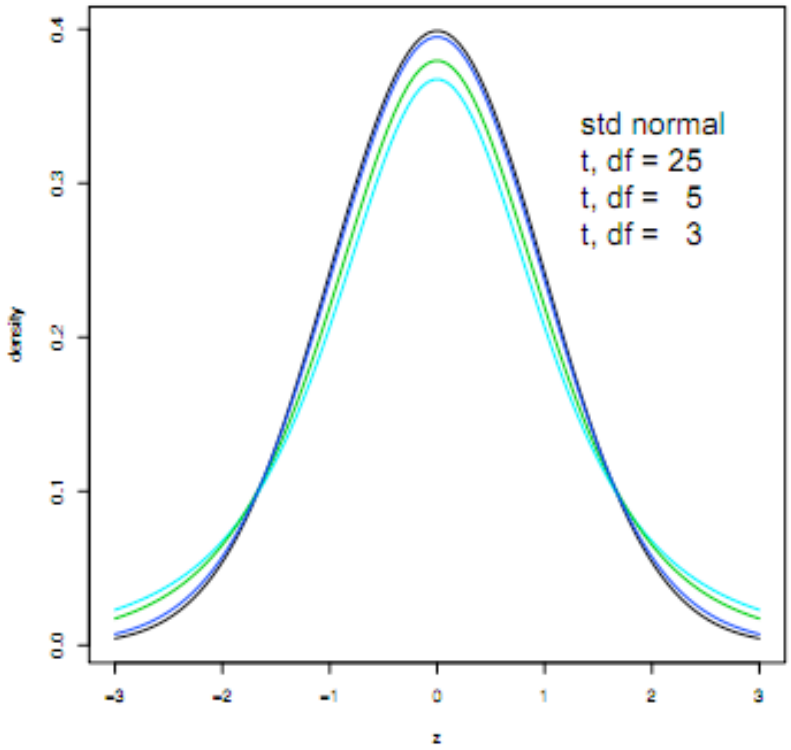
We said there were *n*-1 degrees of freedom in our estimate because 1 was used to compute $\bar{x}$



The *t*-distribution

For small samples (small degrees of freedom) s is quite variable and so we have more spread in the distribution

As we collect larger sample sizes, this variability reduces and we see that the *t*-distribution approaches the standard normal curve



The *t*-distribution

Now, suppose we want a 95% interval using our estimate *s*

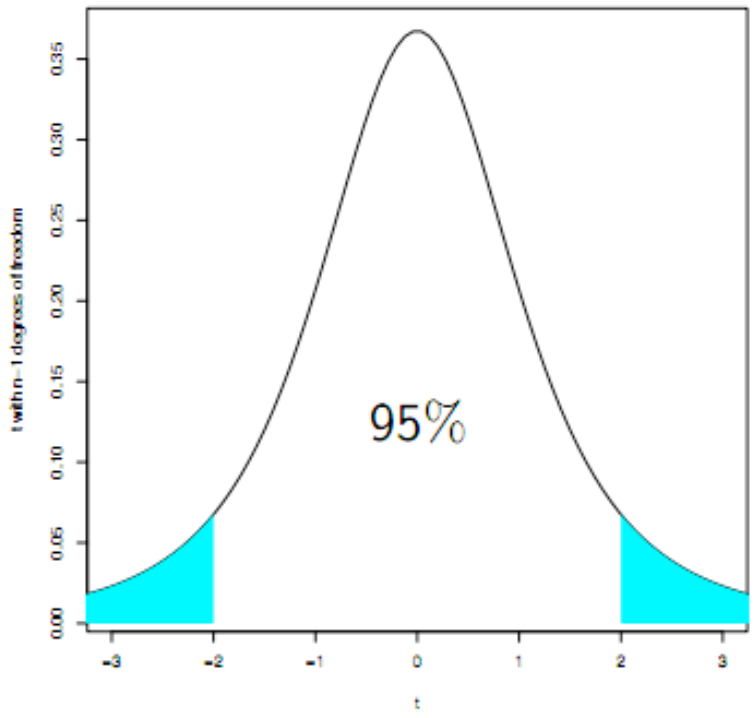
We would would need to find the point *t* such that

$$\text{Prob}\,(-t \leq T \leq t) = 0.95$$

If there are *n* points in our survey, we take the degrees of freedom of the *T* to be *n*-1

We then form the confidence interval

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$



Student's t-distribution

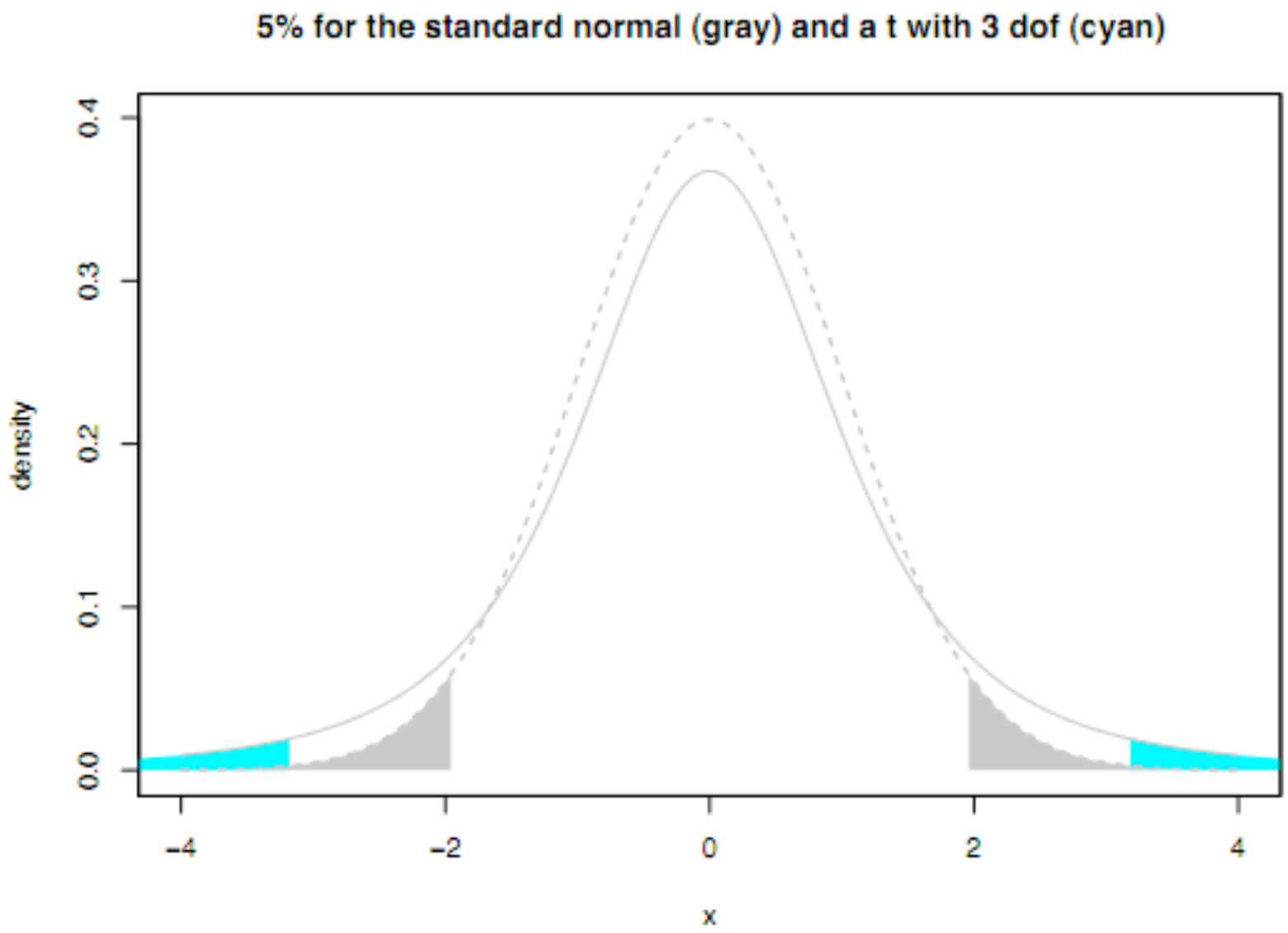
So, what does all this mean? If our data are normally distributed then the *t*-statistic

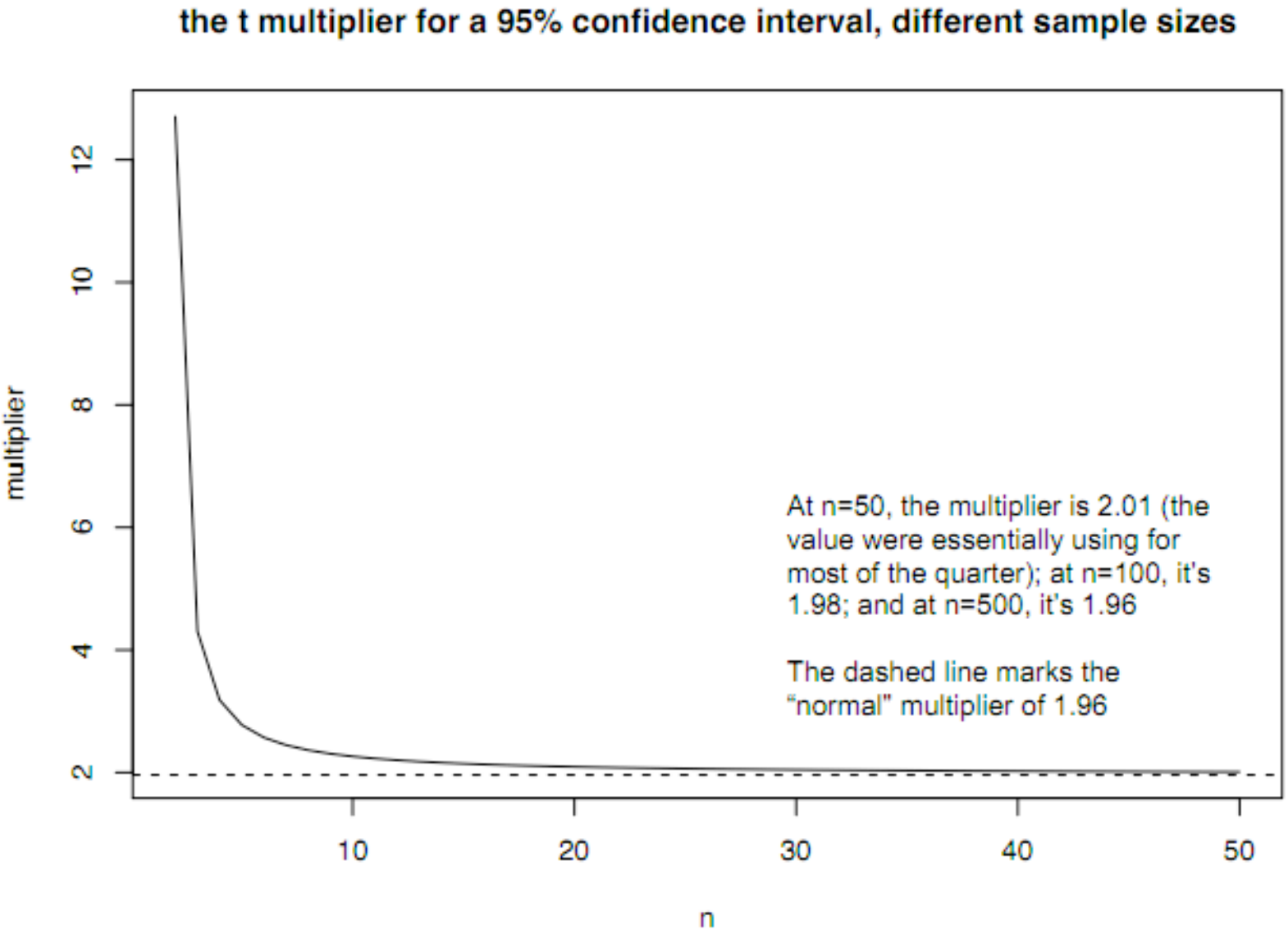
$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a *t*-distribution with n-1 degrees of freedom

As an example, consider Gosset's original simulations with n=4 data points; we would expect 95% of his standardized differences (where 95% refers to repeated experiments) to be within plus or minus `qt(0.975,df=3)` or 3.18 (remember with n=4, we have n-1=3 degrees of freedom)

... and in this case we would use 3.18 instead of 2 (or 1.96) in the multiplier for our confidence interval $\bar{x} \pm 3.18\, s/\sqrt{n}$


5% for the standard normal (gray) and a t with 3 dof (cyan)
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


the t multiplier for a 95% confidence interval, different sample sizes
multiplier
2
4
6
8
10
12
10
20
30
40
50
n
At n=50, the multiplier is 2.01 (the value were essentially using for most of the quarter); at n=100, it's 1.98; and at n=500, it's 1.96
The dashed line marks the "normal" multiplier of 1.96

Student's t-distribution

To sum up; Gosset worked out the sampling distribution of a standardized statistic, **the t-statistic, under the assumption that the data we've observed come from a population with a normal distribution**

Under that assumption, we can derive a confidence interval using quantiles from the t-distribution; **as our sample sizes get large, the effect of estimating σ with s diminishes and we return to the usual normal interval**

A remarkable fact about the t-statistic is that its distribution depends only on the sample size; this is quite a magical fact and one that we will make use of

Summing up

Long before the bootstrap, statisticians were assessing uncertainty in their estimates, often leveraging the Central Limit Effect

In the early 1900s, Gosset (Student) revolutionized statistics by working out the exact sampling distribution under very specific assumptions about the population (that it is normal)

The bootstrap procedure you have learned (and will work with in Lab) agrees with these results when appropriate; the big advantage is that you can estimate the sampling distribution from data and you don't have to hope that the central limit effect has kicked in yet

Because of its generalizability, we will focus primarily on the bootstrap,but we felt it important to see another kind of reasoning...

Epilogue

In 1912, Fisher was an undergraduate studying mathematics at Cambridge University and had just published his first paper “On an Absolute Criterion for Fitting Frequency Curves” in which he introduced the idea of the likelihood function

Noting an n versus n-1 in the formula for the standard deviation used by Gosset , Fisher was encouraged by his advisor to write to Gosset (Gosset was then 36, Fisher 22) and ultimately sent him his proof -- Gosset writes

This prooof, the tutor, made him send me and with some exertion I mastered it, spotted the fallacy (as I believe) and wrote him a letter showing, I hope, an intelligent interest in the matter and incidentally making a blunder. To this he replied with two foolscap pages covered with mathematics of the deepest dye in which he proved, by using n-dimensions that the formula, after all, involved n-1 and, of course, exposing my mistake. I couldn’t understand his stuff and wrote and said I was going to study it when I had time. I actually took it up to the lake with me - and lost it! Now he sends this to me [the mathematical proof of Student’s distribution]. It seemed to me that if it’s all right, perhaps you might like to put the proof in a note. It’s so nice and mathematical that it might appeal to some people.

... the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution... : since some law of distribution must be assumed, it is better to work with a curve whose area and ordinates are tabled and whose properties are well known. This assumption is accordingly made in the present paper so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error.

- Student's 1908 paper

What I should like you to do is to find a solution for some other population than a normal one. It seems to me you might assume some sort of an equation for the frequency distribution of x which could lend itself to treatment besides the Gaussian..

- Letter to Fisher from Gosset

I have never known difficulty to arise in biological work from imperfect normality of the variation, often though I have examined data for this particular cause of difficulty; nor is there, I believe, any case to the contrary in the literature.

- Letter to Gosset from Fisher

Fisher is only talking through his hat when he talks of his experience; it isn't so very extensive and I bet he hasn't often put the matter to the test; how could he?

-Gosset writing to E. Pearson

The existence of these random numbers [Tippet's tables] opened out the possibility scarcely dreamed of before, of carrying out a great variety of experimental programmes, particularly answering questions in considerable depth and breadth the kind of questions about robustness of the 'normal theory' tests based on z (or t) raised by Gosset in his letter to me of 11 May 1926. This programme I started in 1927...

- From E. Pearson's memoir