Probability distributions and confidence limits

20.109

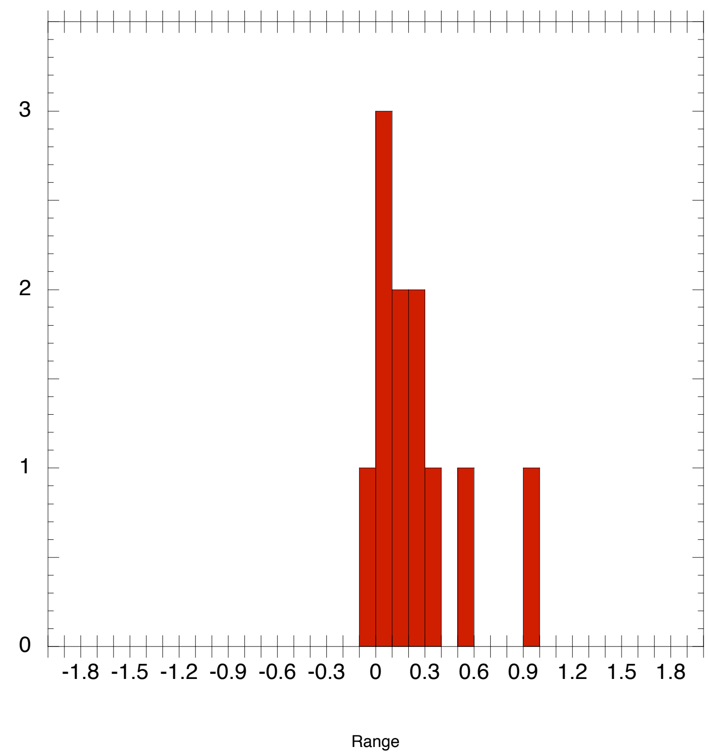
- Probability distributions
  - Gaussian or normal
  - Poisson
- Quantifying uncertainty about parameters
  - confidence limits

Random numbers

- Numbers that are not precisely predictable
- In repeated trials, the distribution of outcomes will map out a probability distribution
- Common probability distributions:
  - Gaussian or normal
  - Poisson

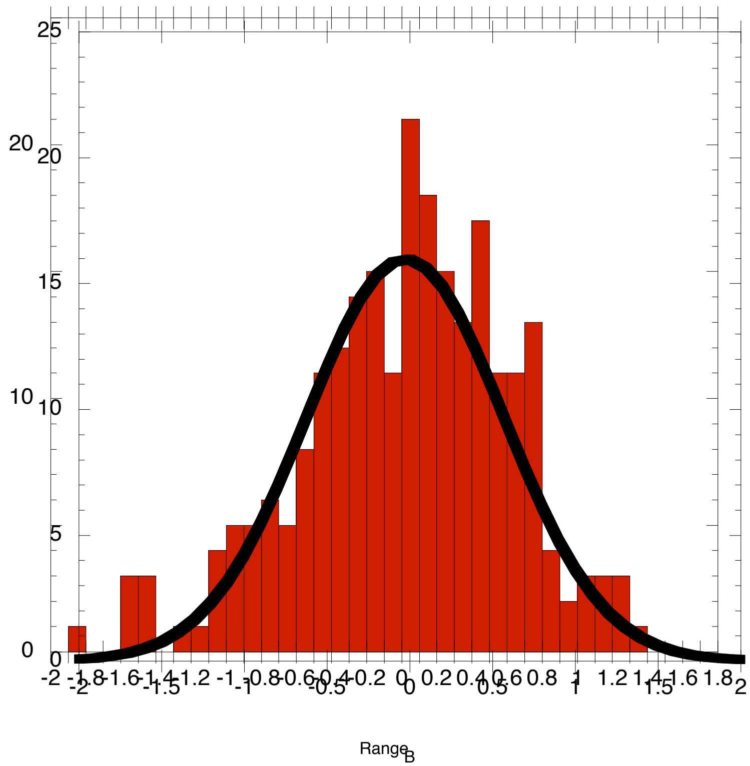
Example: MS vs. sequence MW

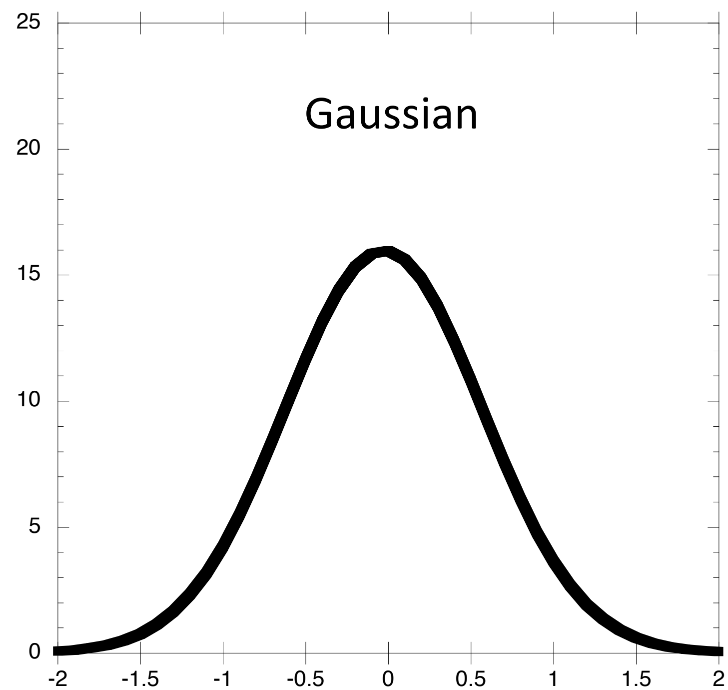
Discrepancies:
-0.1
0.0
0.0
0.1
0.3
0.9
0.5
0.2
0.2
0.1
0



11 experiments

Example cont'd

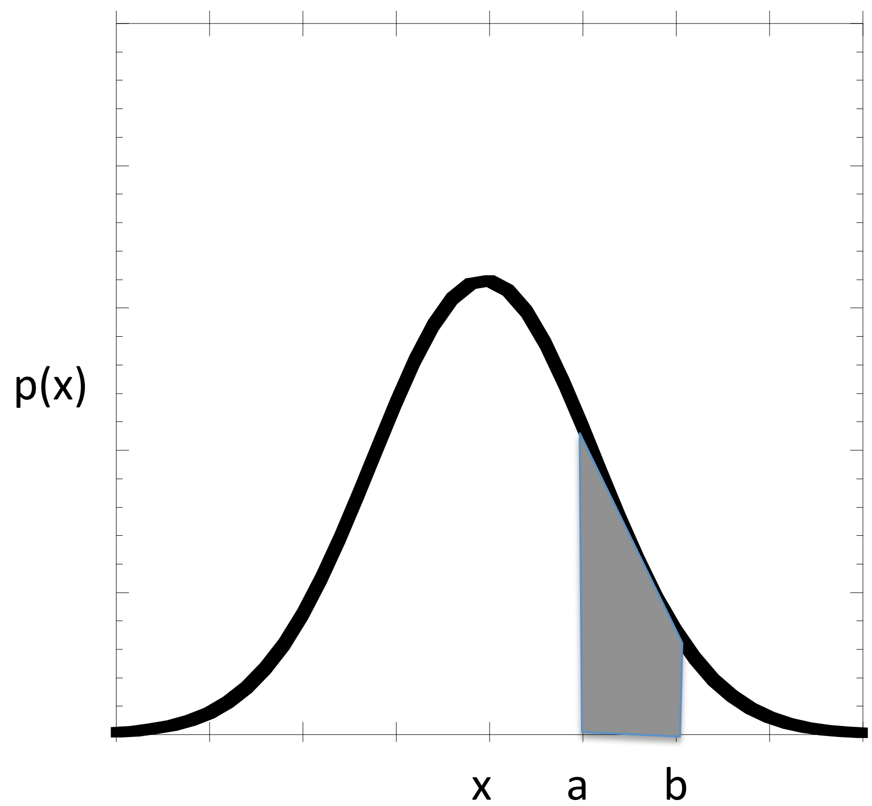




239 data points
Average = -0.027
Standard deviation = 0.597

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Probability density function p(x)



x is a random number

Normalized

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Probability that

$$a < x < b$$

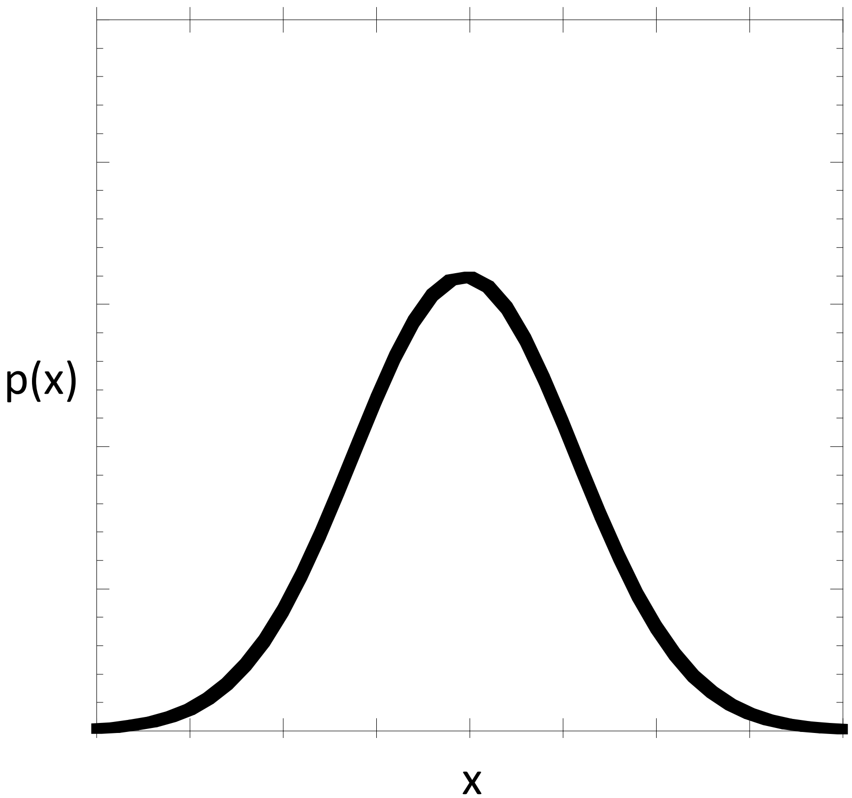
is

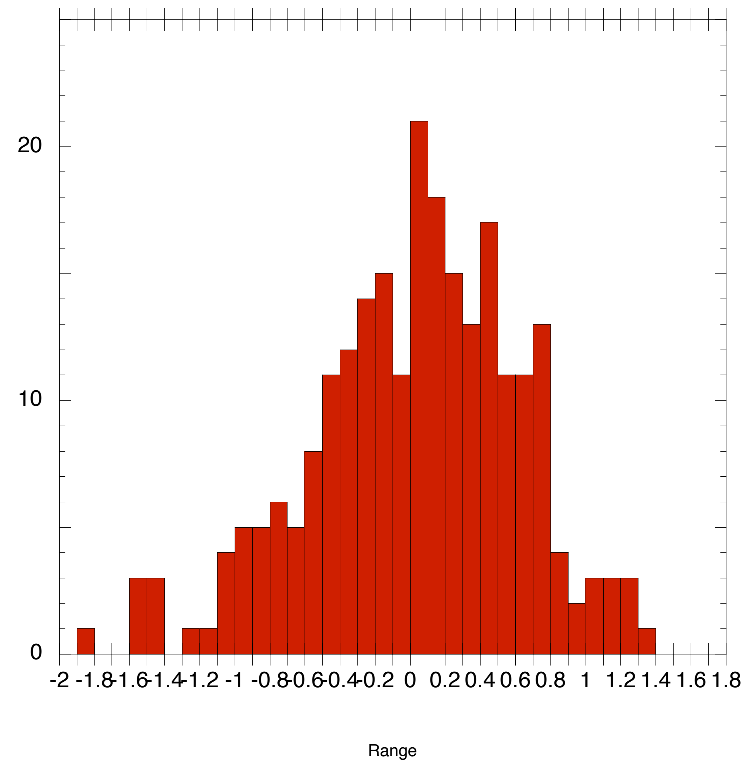
$$\int_{a}^{b} p(x)dx$$

Types of probability density functions

- Uniform
- Gaussian or normal
- Poisson
- Binomial
- Geometric
- ……

Truth vs. sample estimation





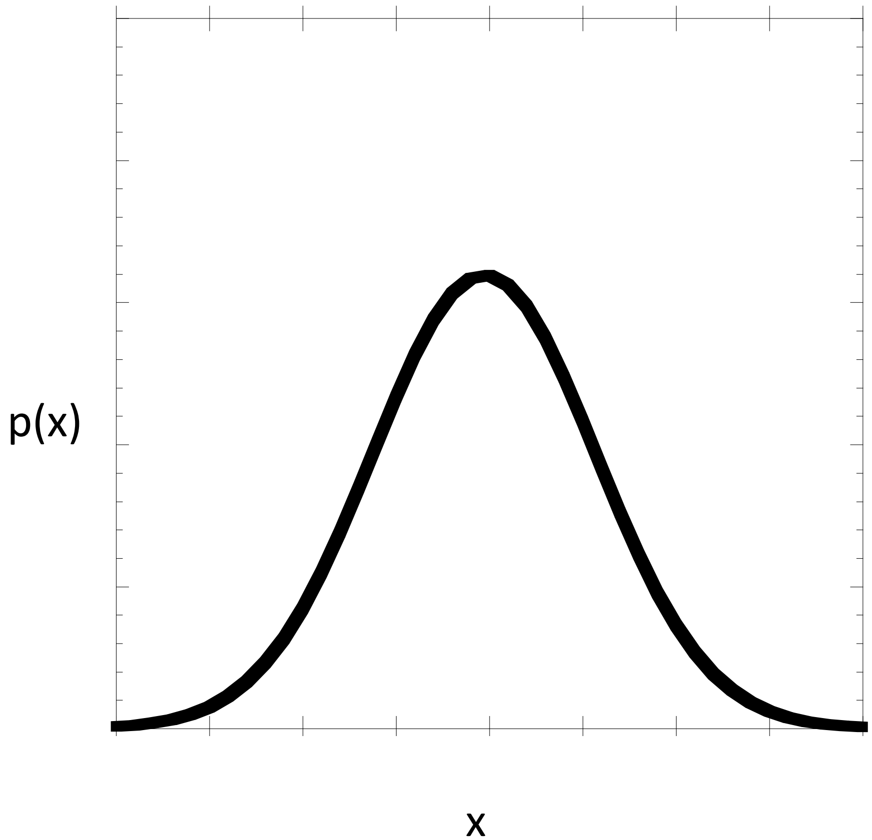
$$\mu = \int_{-\infty}^{\infty} x p(x) dx$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Gaussian or normal



The sum of a large number of independent random variables is normally distributed

Also the solution to a 1-D random walk/Brownian motion/ diffusion problem

Many measurements follow this distribution (e.g. mass spec example previous)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Poisson distribution

Discrete events, counted in sample volumes or times

Assumptions:

1. In a small enough increment in space or time, only zero or one event will occur.
2. Events in each increment of space or time are independent of events in every other increment.
3. The probability of success is proportional to the size of the increment

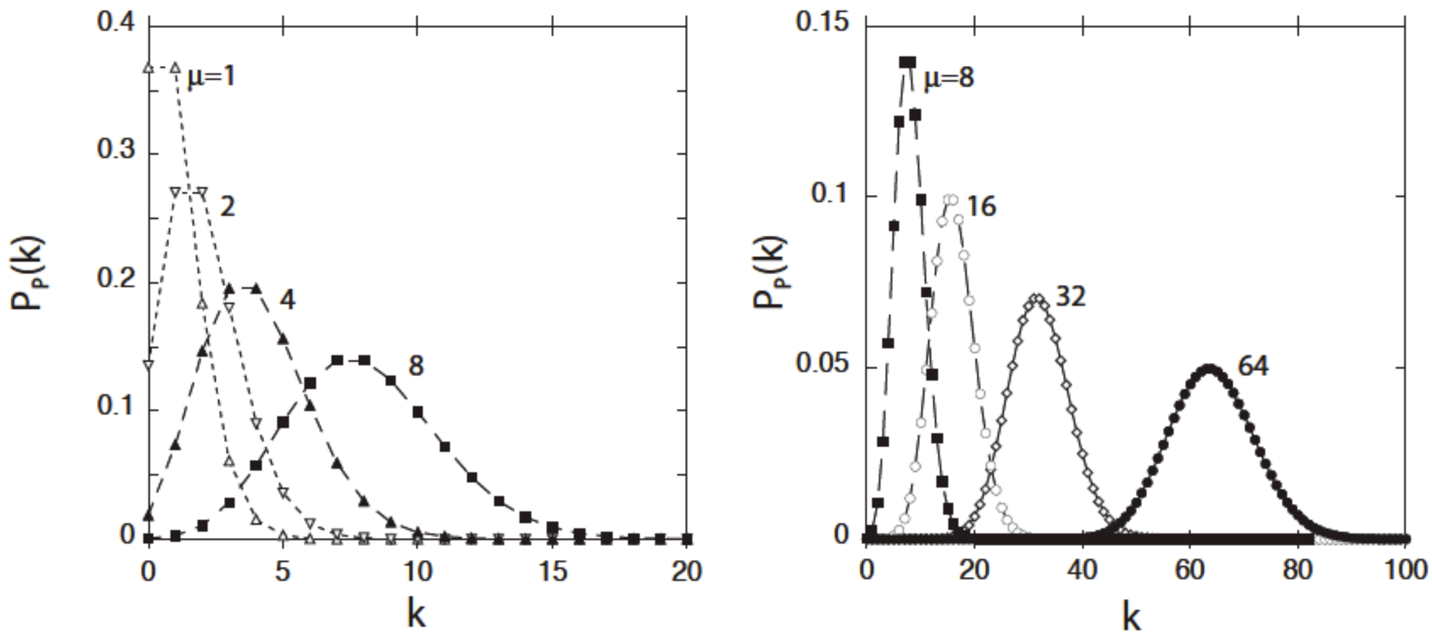
Examples:

Tics of a geiger counter in a fixed time interval
Raisins in a spoonful of pudding
Colonies on a Petri dish

$$p(k) = \frac{\mu^k}{k!} e^{-\mu}$$

Where μ = average

Poisson distribution (cont'd)



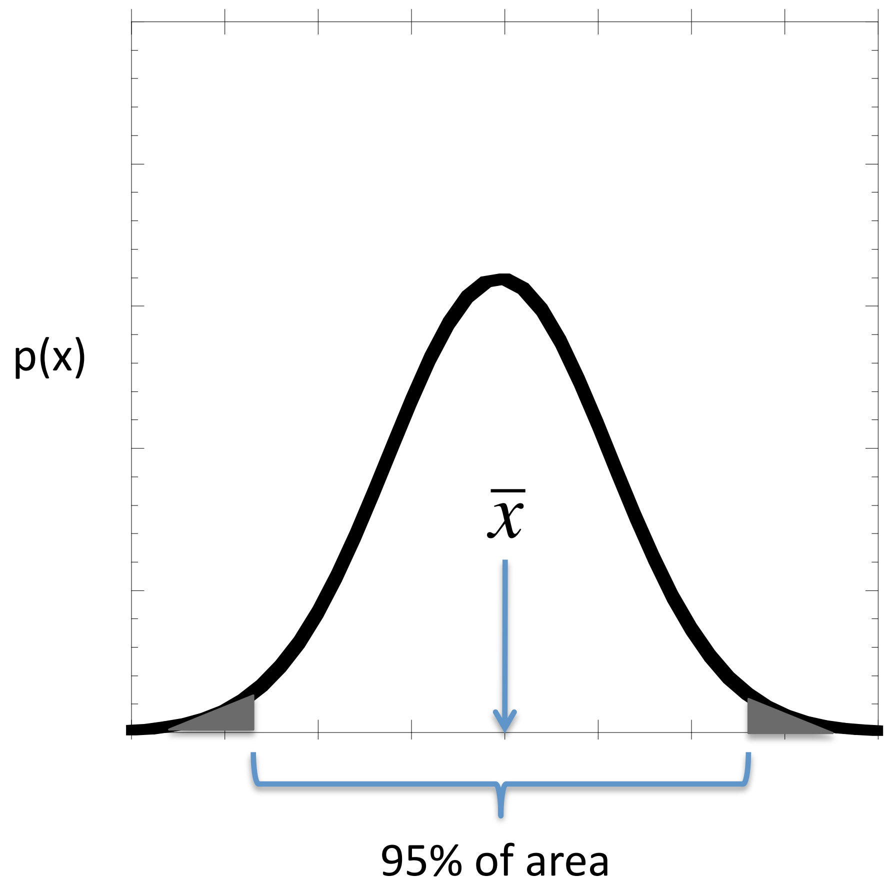
Poisson approaches Gaussian form as μ increases

$$\sigma^2 = \mu$$

Variance = mean

95% confidence interval of an estimate

A range such that 95% of replicate estimates would be within it



Common but less rigorous practice

$\bar{x} \pm s$ is often reported

Which in a normal distribution encompasses 66% of the area

However, both $\bar{x}$ and s are only estimates

So in effect we're unsure about how unsure we are!

95% Confidence interval for a normally distributed variable

$$\bar{x} - \frac{t_{0.025} s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{0.025} s}{\sqrt{n}}$$

# data points	$t_{0.025}$
2	12.706
3	4.303
4	3.182
5	2.776
10	2.262
20	2.093
30	2.045
50	2.010
100	1.984

Increasingly accurate estimate of σ

Note: Uncertainty decreases proportionally to $\frac{1}{\sqrt{n}}$

So take more data!

Example

3 measurements of absorbance at 600 nm: 0.110, 0.115, 0.113

95% confidence limit?

Soln:

$$\bar{x} = 0.113, s = 0.0025$$

$$\bar{x} - \frac{t_{0.025}s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{0.025}s}{\sqrt{n}}$$

$$0.113 - \frac{4.303(0.0025)}{\sqrt{3}} < \mu < .113 + \frac{4.303(0.0025)}{\sqrt{3}}$$

$$0.107 < \mu < 0.119$$

95% confidence interval for a Poisson variable

Could actually sum up the probabilities for 1, 2, etc. to exactly find the interval; or look it up in a table

Alternative approximation:

$$\left(\frac{z_{0.025}}{2} - \sqrt{\bar{x}}\right)^2 < \mu < \left(\frac{z_{0.025}}{2} + \sqrt{\bar{x}+1}\right)^2$$

$$z_{0.025} = 1.96$$

Note: interval is not symmetric but approaches it at larger $\bar{x}$

European Journal of Epidemiology, Vol. 20, No. 3 (2005), pp. 213-216

Example

47 colonies on a plate from 20 microliters plated. 95% confidence interval?

Soln:

$$\left(\frac{z_{0.025}}{2}-\sqrt{\bar{x}}\right)^2<\mu<\left(\frac{z_{0.025}}{2}+\sqrt{\bar{x}+1}\right)^2$$

$$\left(\frac{1.96}{2}-\sqrt{47}\right)^2<\mu<\left(\frac{1.96}{2}+\sqrt{47+1}\right)^2$$

$$34.5<\mu<62.5$$

Confidence limit for a fraction

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\left(1-\hat{p}\right)/n}$$

$$z_{0.025} = 1.96$$

$$\hat{p} = \frac{n_{success} + 2}{n + 4}$$

American Statistician 52:119, '98

Example

47 colonies on selective medium, 83 colonies on nonselective.
95% confidence limit on plasmid-containing fraction?
Soln:

$$\hat{p} = \frac{n_{success} + 2}{n + 4}$$

$$\hat{p} = \frac{47 + 2}{83 + 4} = 0.56$$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1 - \hat{p})/n}$$

$$0.56 \pm 1.96\sqrt{0.56(1 - 0.56)/83}$$

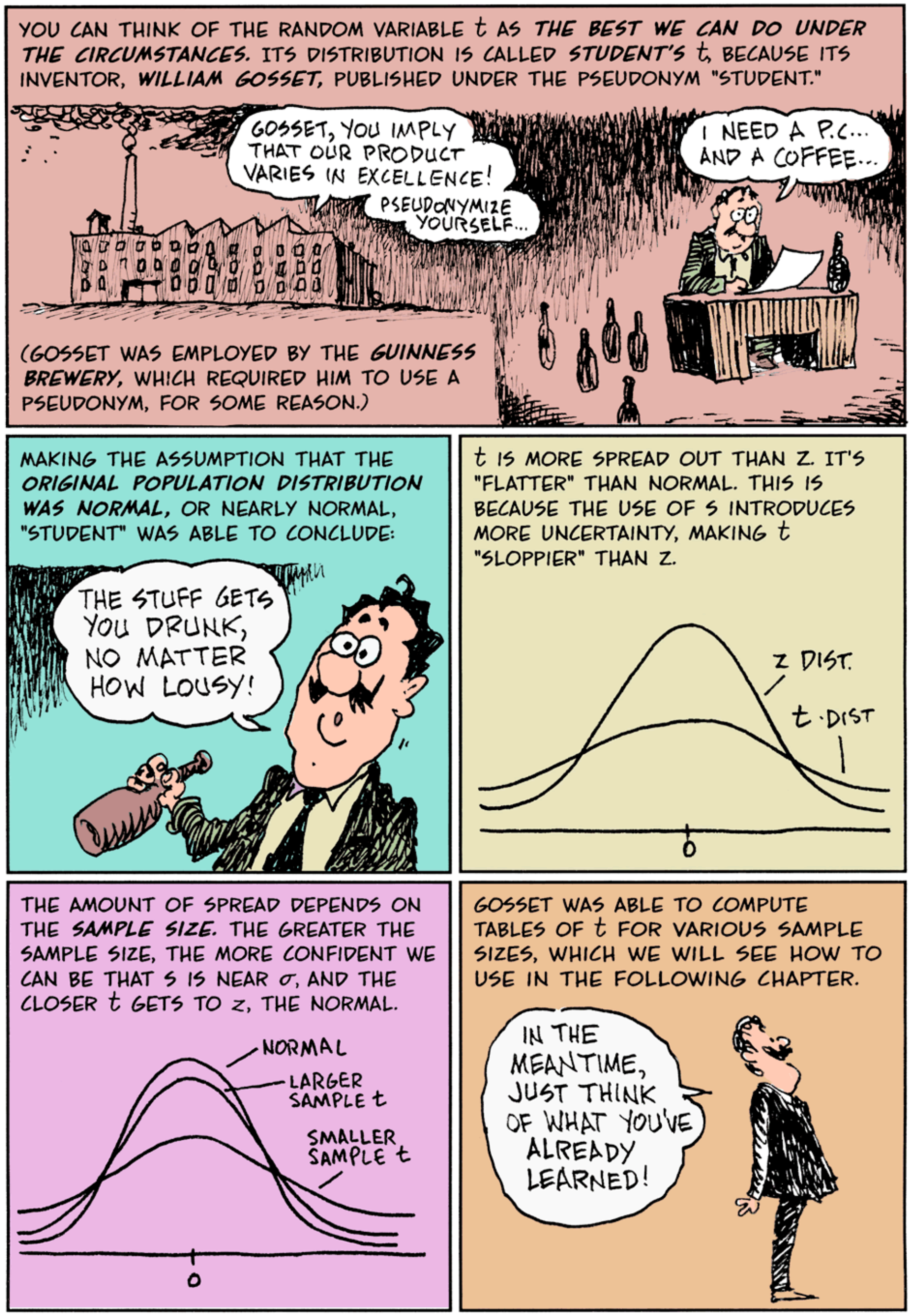
$$0.56 \pm 0.11$$

Where t tests come from

Which barley variety

Makes better stout?

(Danish Archer)

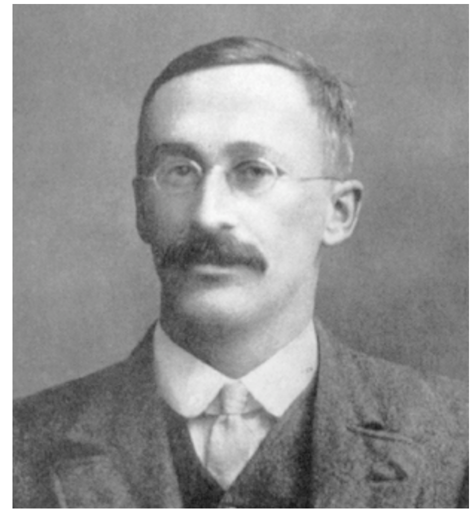

YOU CAN THINK OF THE RANDOM VARIABLE t AS THE BEST WE CAN DO UNDER THE CIRCUMSTANCES. ITS DISTRIBUTION IS CALLED STUDENT'S t, BECAUSE ITS INVENTOR, WILLIAM GOSSET, PUBLISHED UNDER THE PSEUDONYM "STUDENT."
GOSSET, YOU IMPLY THAT OUR PRODUCT VARIES IN EXCELLENCE!
PSEUDONYMIZE YOURSELF...
I NEED A P.C... AND A COFFEE...
(GOSSET WAS EMPLOYED BY THE GUINNESS BREWERY, WHICH REQUIRED HIM TO USE A PSEUDONYM, FOR SOME REASON.)
MAKING THE ASSUMPTION THAT THE ORIGINAL POPULATION DISTRIBUTION WAS NORMAL, OR NEARLY NORMAL, "STUDENT" WAS ABLE TO CONCLUDE:
THE STUFF GETS YOU DRUNK, NO MATTER HOW LOUSY!
t IS MORE SPREAD OUT THAN Z. IT'S "FLATTER" THAN NORMAL. THIS IS BECAUSE THE USE OF s INTRODUCES MORE UNCERTAINTY, MAKING t "SLOPPIER" THAN Z.
Z DIST.
t-DIST
0
THE AMOUNT OF SPREAD DEPENDS ON THE SAMPLE SIZE. THE GREATER THE SAMPLE SIZE, THE MORE CONFIDENT WE CAN BE THAT s IS NEAR σ, AND THE CLOSER t GETS TO z, THE NORMAL.
NORMAL
LARGER SAMPLE t
SMALLER SAMPLE t
0
GOSSET WAS ABLE TO COMPUTE TABLES OF t FOR VARIOUS SAMPLE SIZES, WHICH WE WILL SEE HOW TO USE IN THE FOLLOWING CHAPTER.
IN THE MEANTIME, JUST THINK OF WHAT YOU'VE ALREADY LEARNED!

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.



"It may seem strange that reasoning of this nature had not been more widely made use of, but this is due, first, to the popular dread of mathematics."
W.S. Gossett