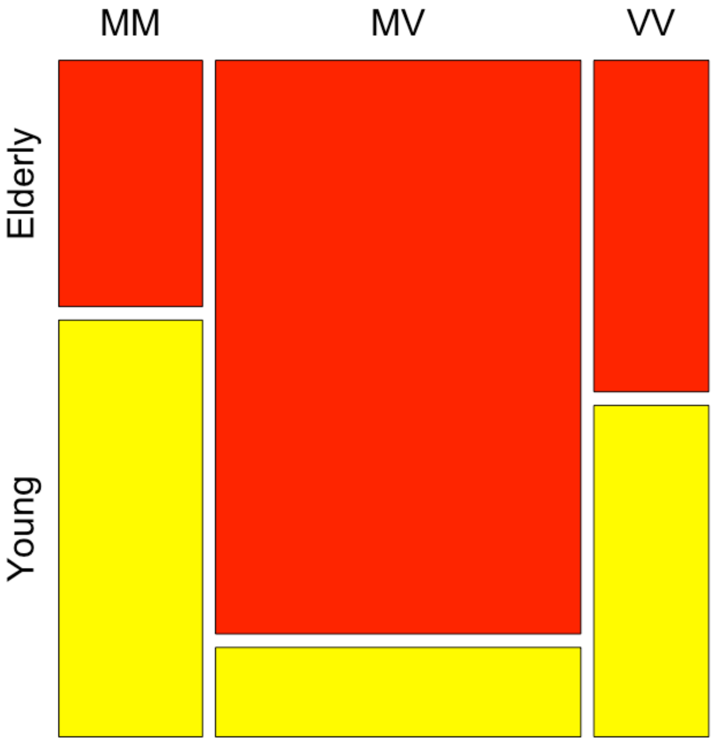
** On analyzing Contingency Tables

** The normal distribution

Week 08

Kuru dataset



```
kuru=
matrix(data=c(13,77,14,22,12
,14), nrow =2, ncol =3,
byrow = T)
rownames(kuru)=c("Elderly",
"Young")
colnames(kuru) = c("MM",
"MV", "VV")
mosaicplot(t(kuru), main="",
col=c("red","yellow"),
cex=2)
```

Test for contigency tables

Chi square test (9.4)

Df = (r − 1) * (c-1)

Just another G o F test

G-test (9.6, see likelihood)

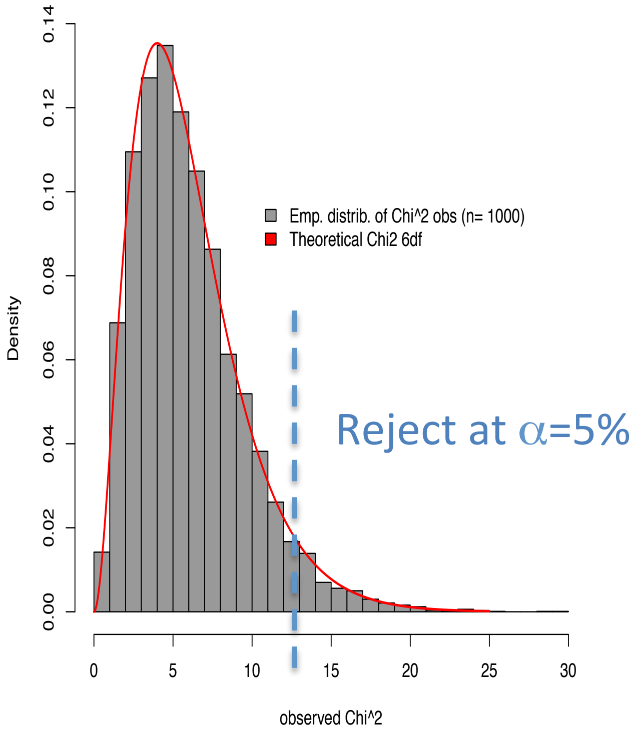
Fisher's exact test (9.5)

Enumerate all tables that are more "extreme" and sums up their probabilities… actuallt HARD to do … works great for small counts

Remember …. A well behaved test

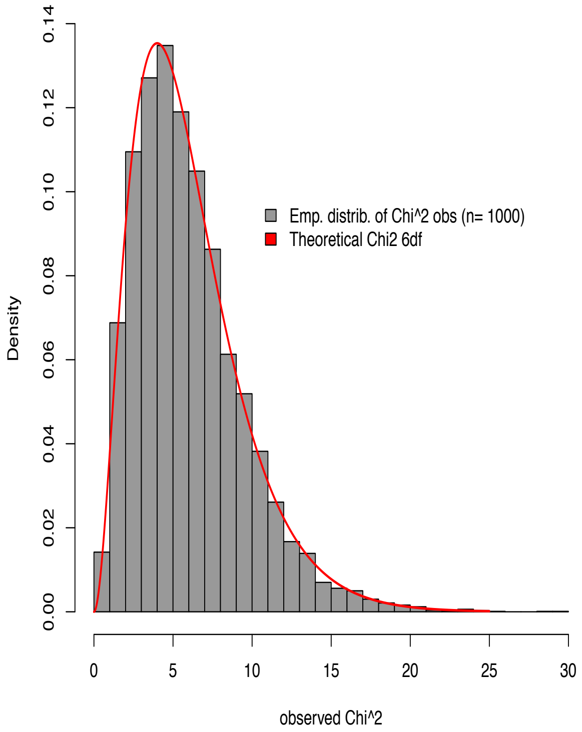
Test statistic from Data under H0 is well approximated by the limiting distribution

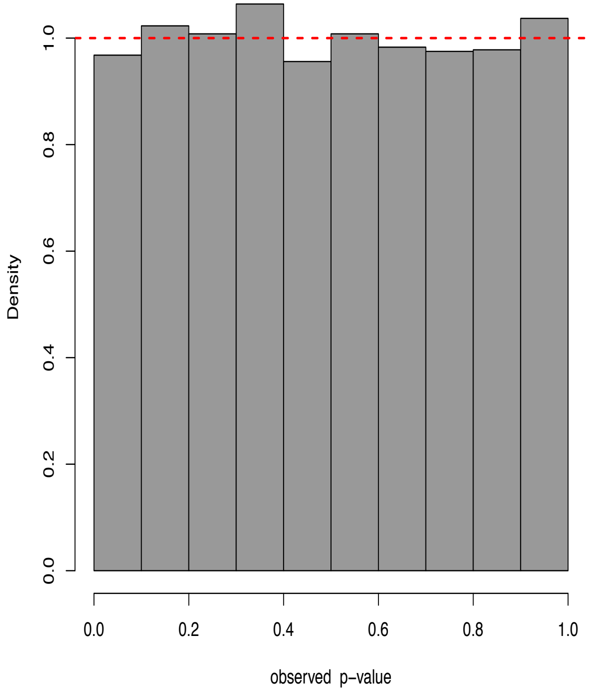
→ P-values are …?



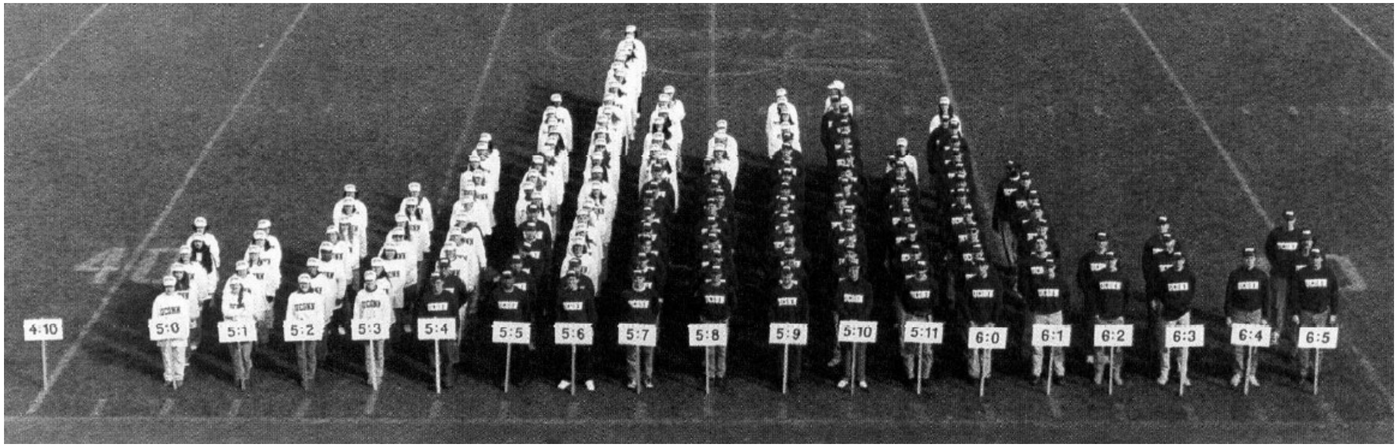
A well behaved test

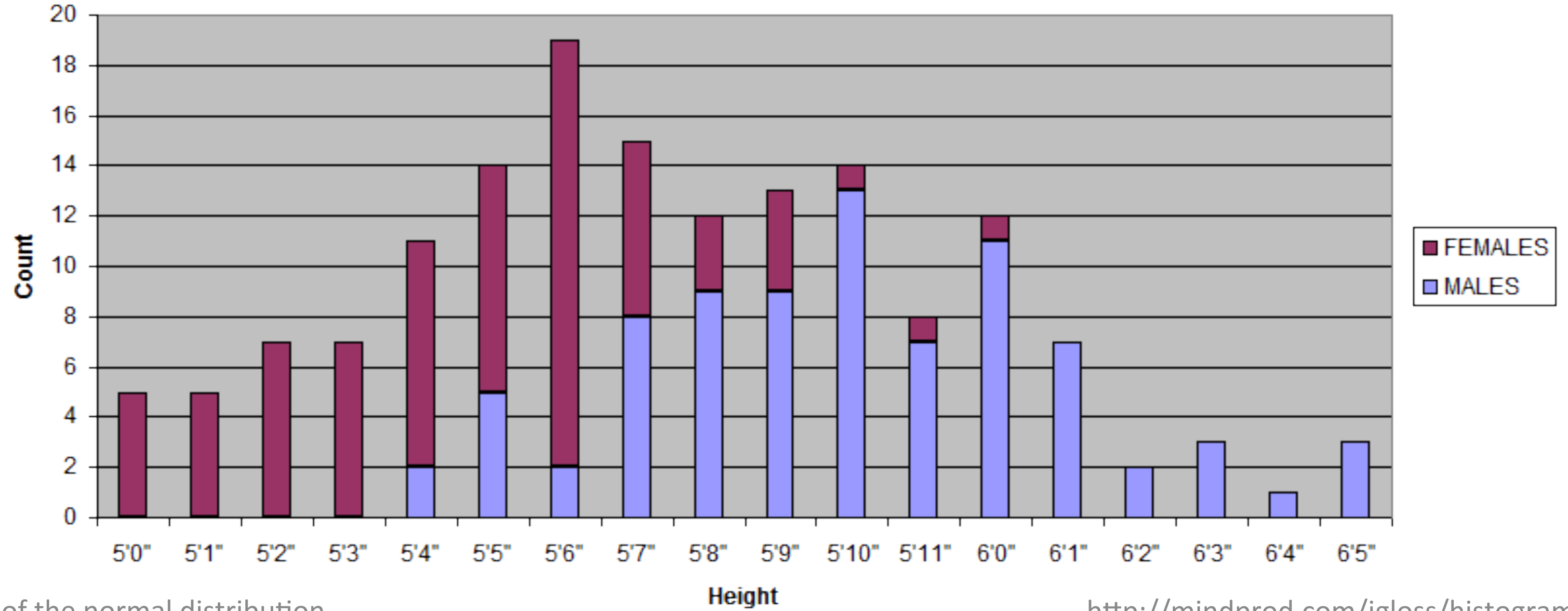
Data simulated under H0 is well approximated by limiting distribution ← → **P-values are uniform in [0,1]**





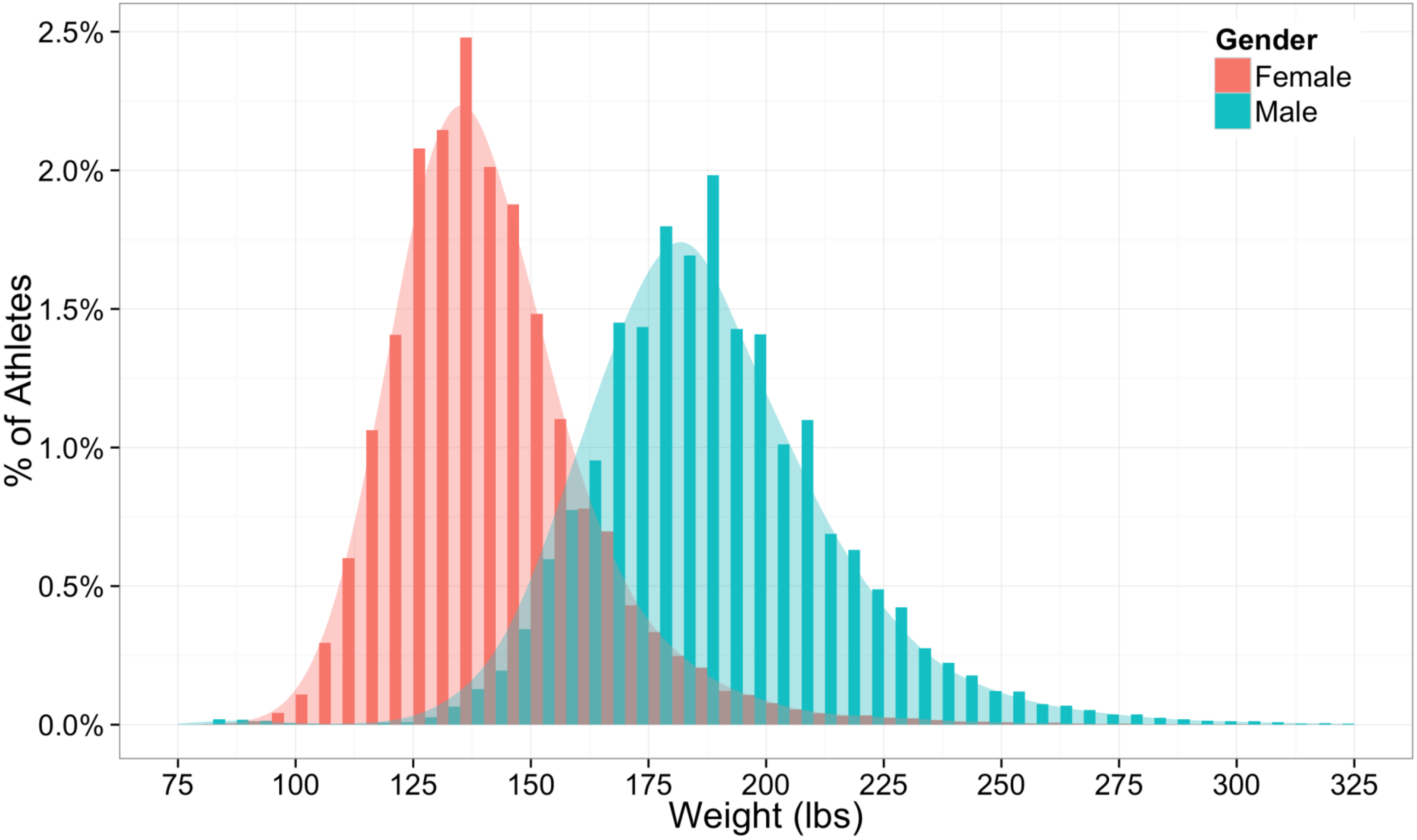
Height of college students



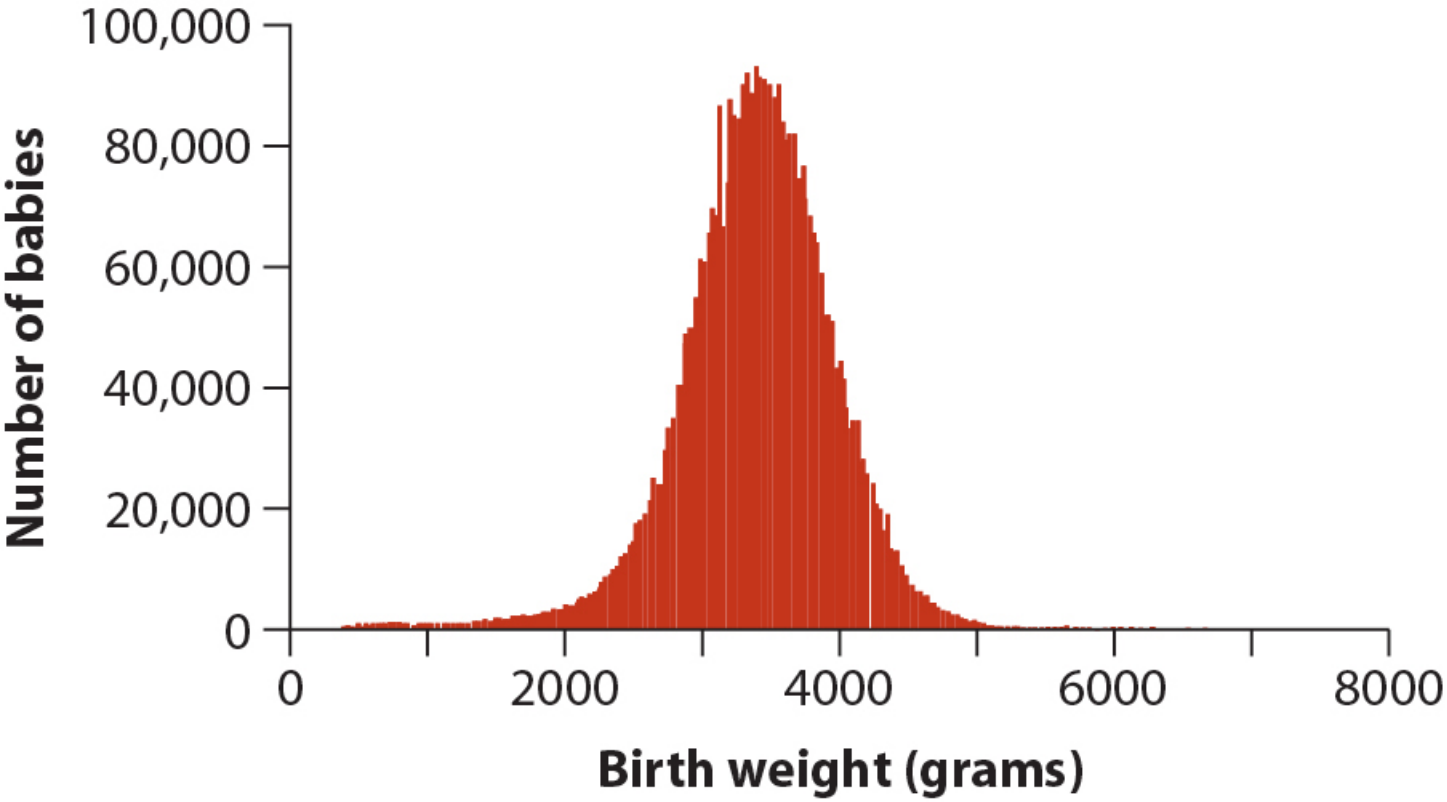


http://mindprod.com/jgloss/histogram.html

Weight of women vs men

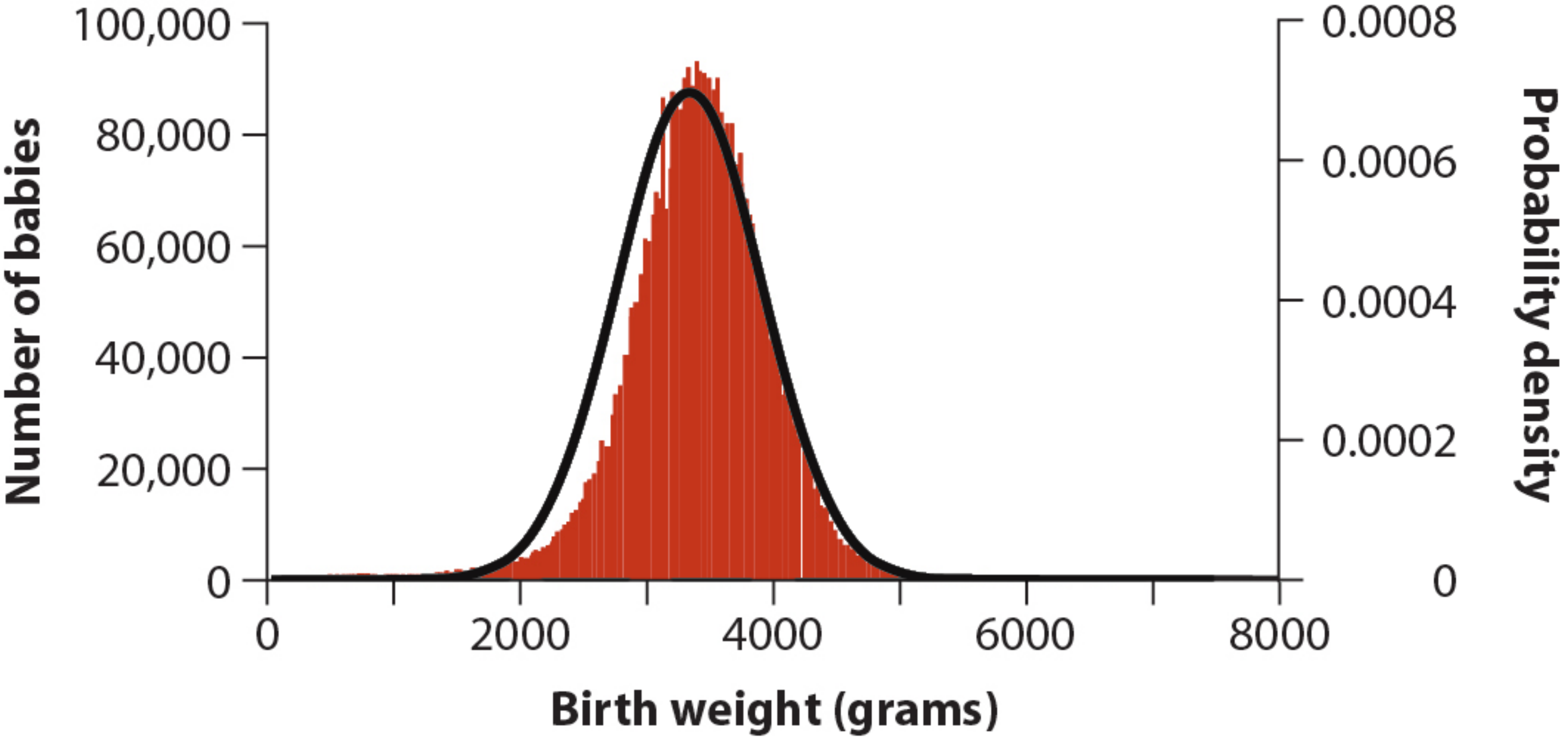


Normal data



Number of babies
100,000
80,000
60,000
40,000
20,000
0
0
2000
4000
6000
8000
Birth weight (grams)

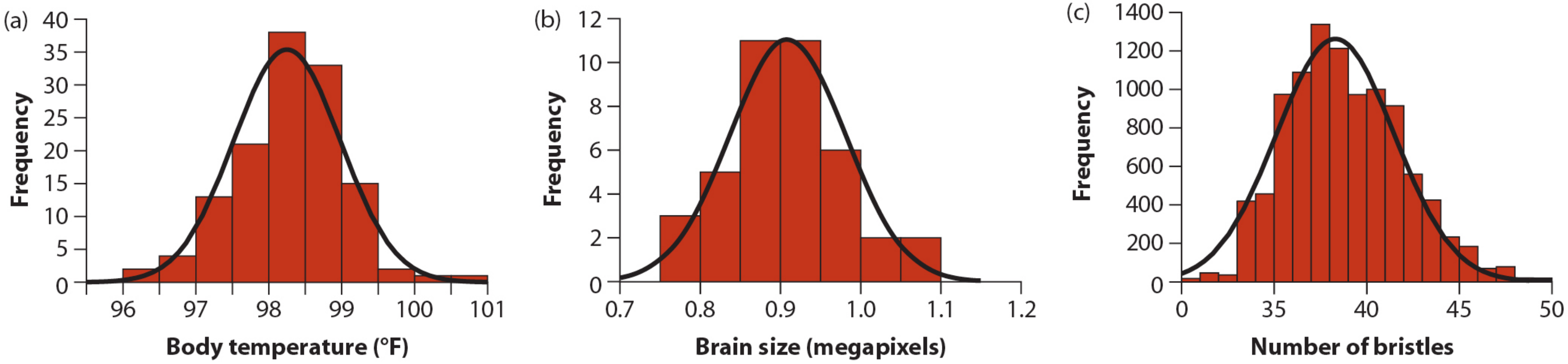
Normal distribution



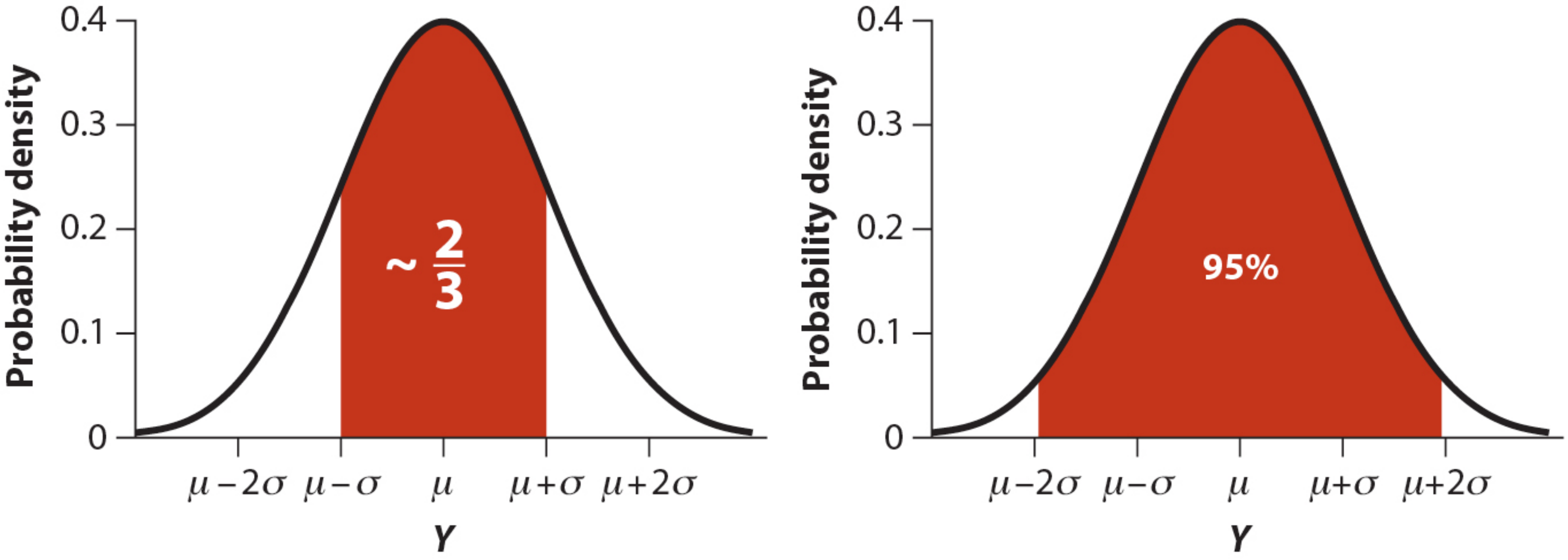
$f(Y \mid \mu, \sigma)= 1/\sigma\ \sqrt{2}\pi\ \ e\uparrow-(Y-\mu)\uparrow 2\ /2\sigma\uparrow 2$

Common in nature

- Many biological measures are (almost) normally distributed
  - body temperature
  - brain size
  - number of bristles (discrete)
  - height
  - weight
  - BMI



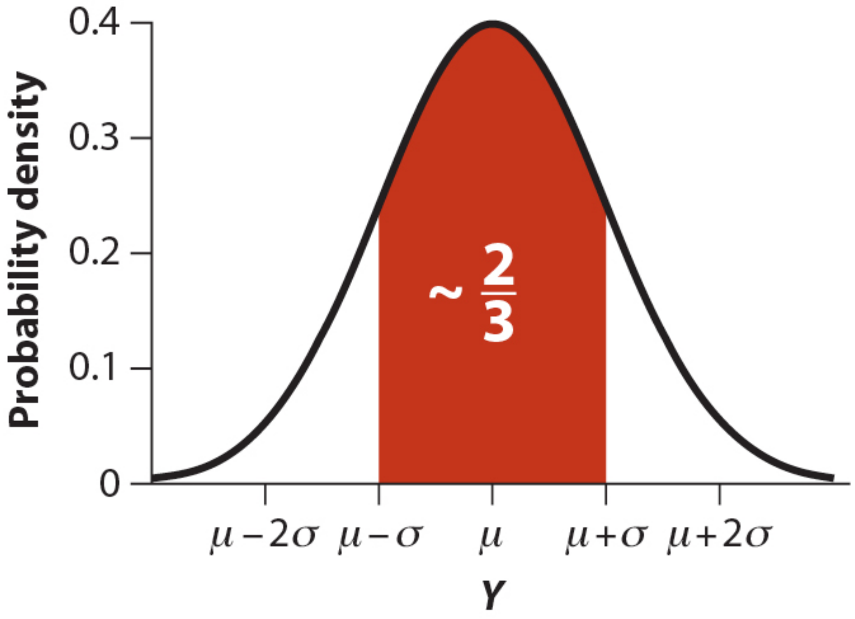
Properties

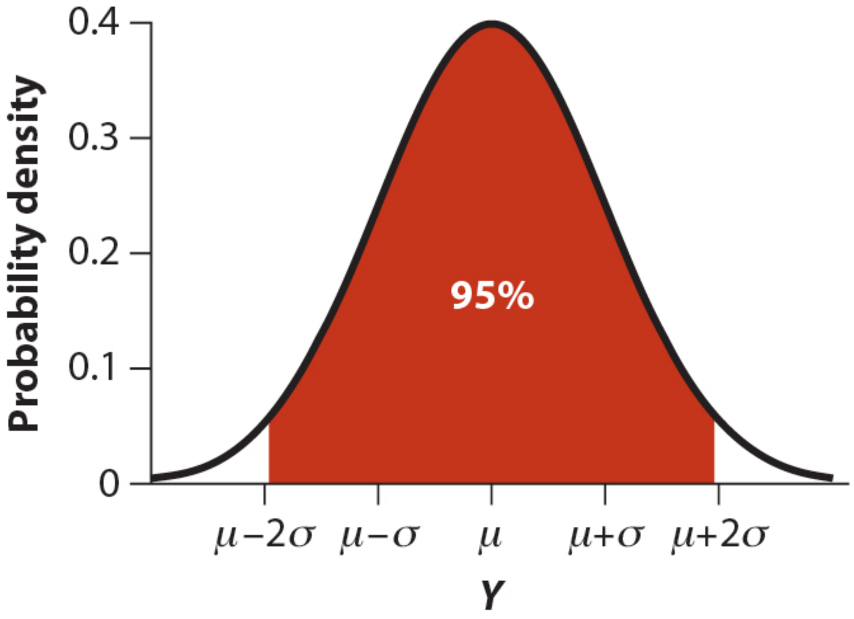


$\Pr[\mu - \sigma \leq x \leq \mu - \sigma] \approx 0.683$

$\Pr[\mu - 2\sigma \leq x \leq \mu - 2\sigma] \approx 0.955$

Properties



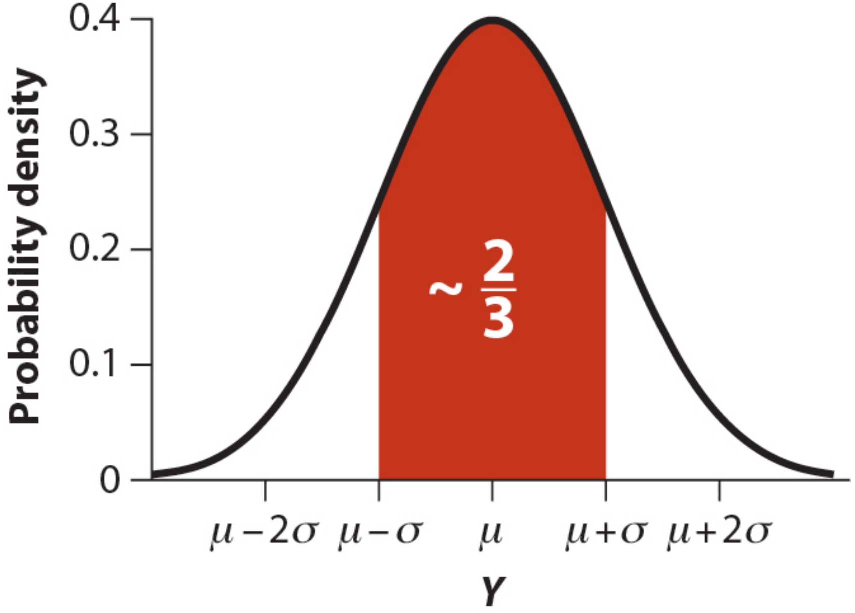


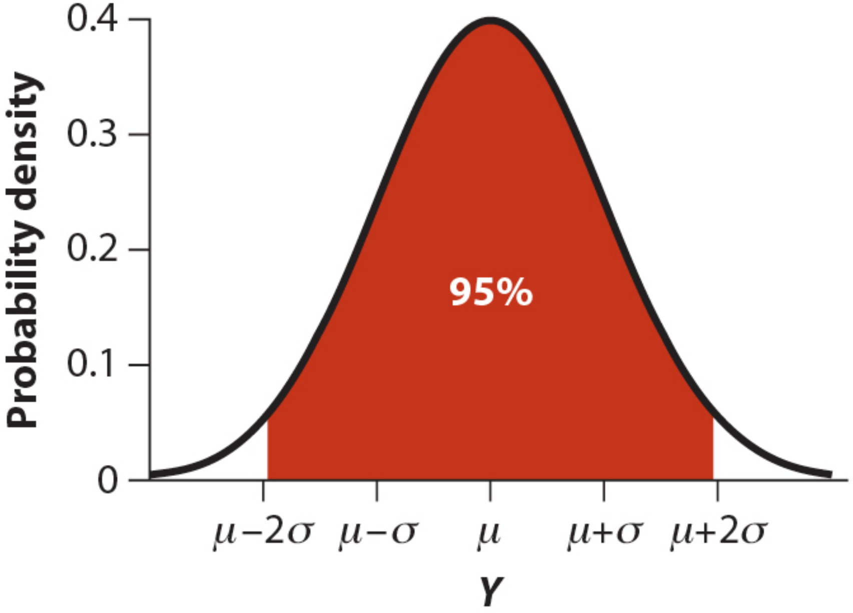
$\mathrm{Pr}[\mu - \quad \sigma \leq x \leq \mu - \quad \sigma] \approx 0.683$

$\mathrm{Pr}[\mu - \quad 2\sigma \leq x \leq \mu - \quad 2\sigma] \approx 0.955$

Approximate 95% CI (chapter 4): $[Y - 2SE{\downarrow}Y \ , Y + 2SE{\downarrow}Y \ ]$

Properties





Pr$[\mu - \quad \sigma \leq x \leq \mu - \quad \sigma] \approx$ 0.683

Pr$[\mu - \quad 2\sigma \leq x \leq \mu - \quad 2\sigma] \approx$ 0.955

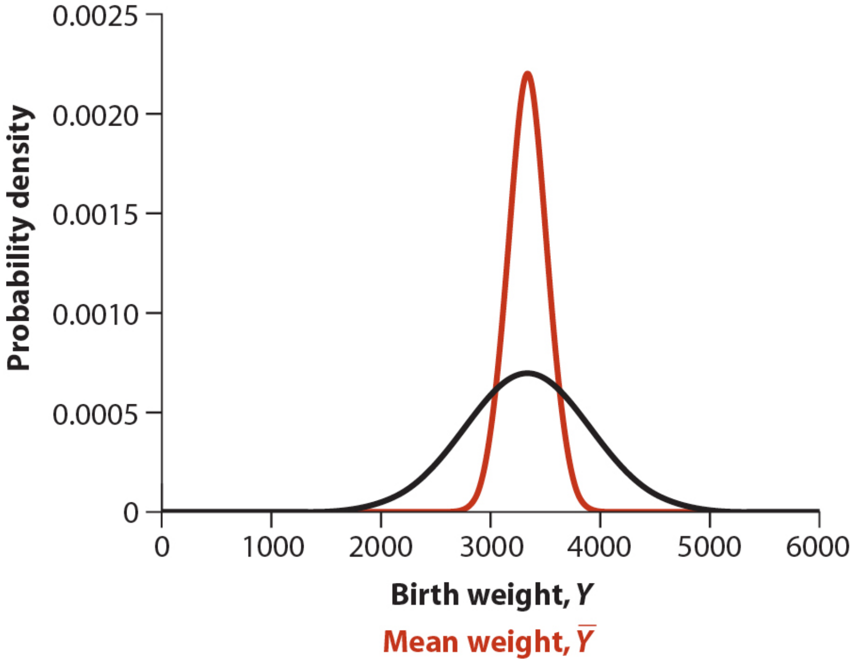
Approximate 95% CI (chapter 4): $[Y - 2SE{\downarrow}Y\ , Y + 2SE{\downarrow}Y\ ]$

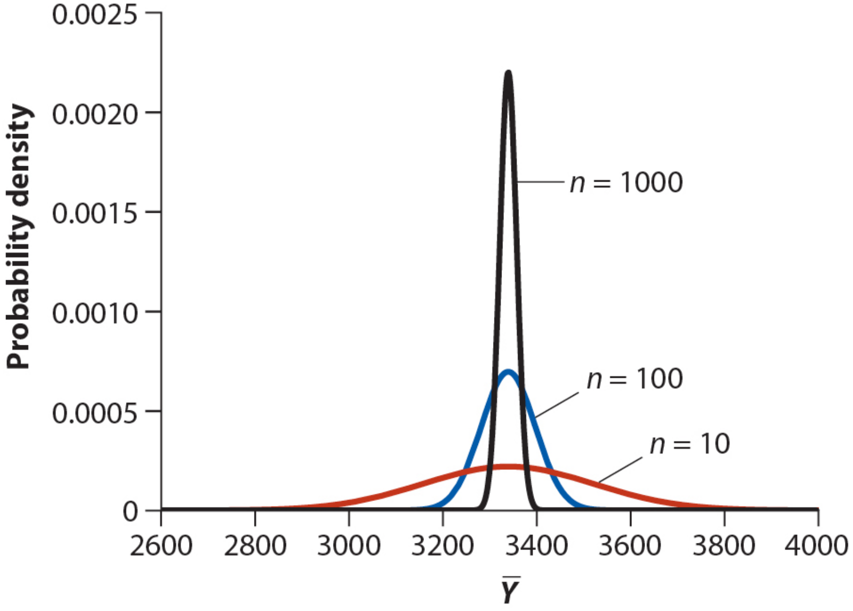
Pr$[\mu - 1.96\sigma \leq x \leq \mu - 1.96\sigma] \approx$ 0.950

Normal distribution in R

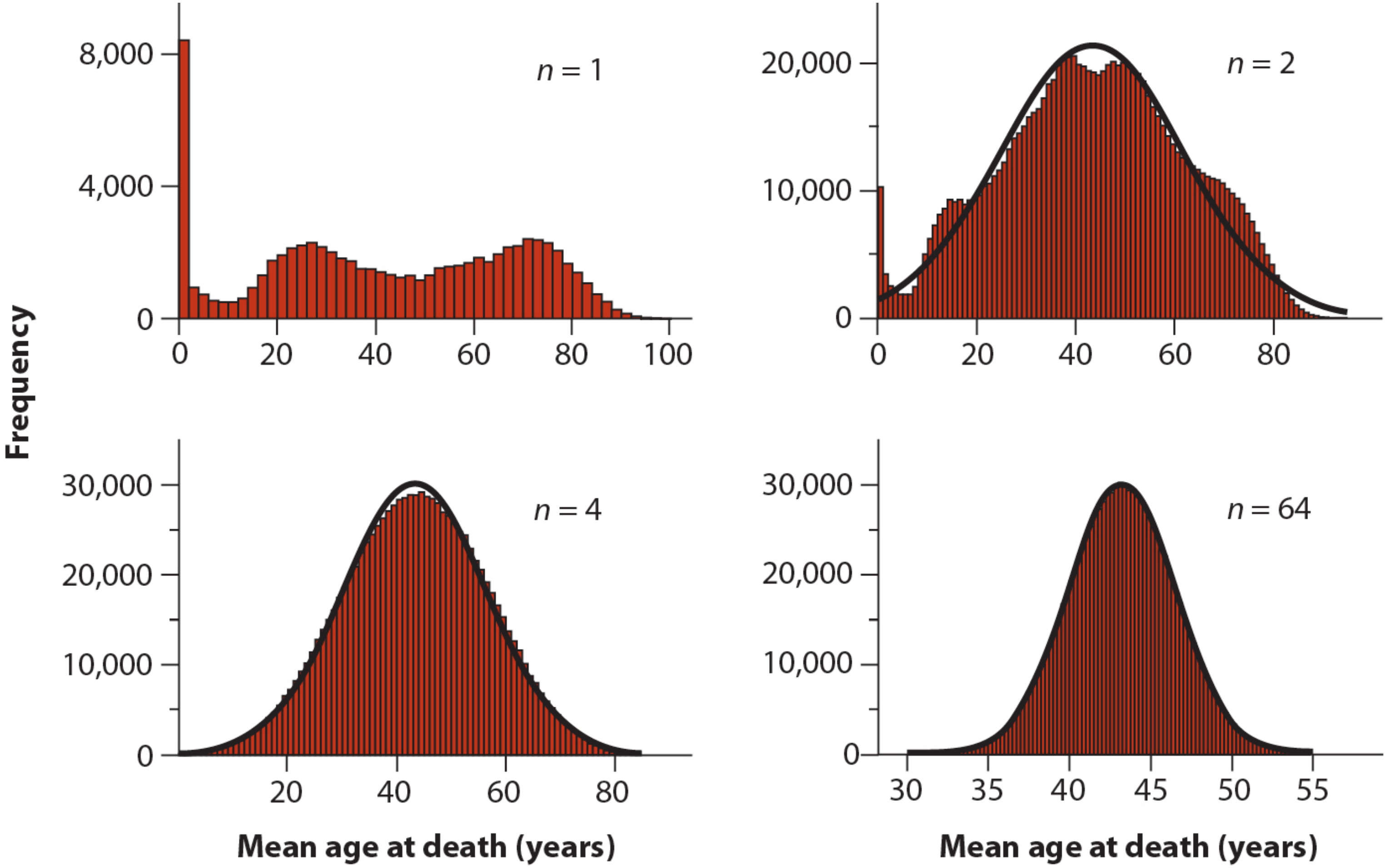
- rnorm pnorm dnorm

The distribution of sample means





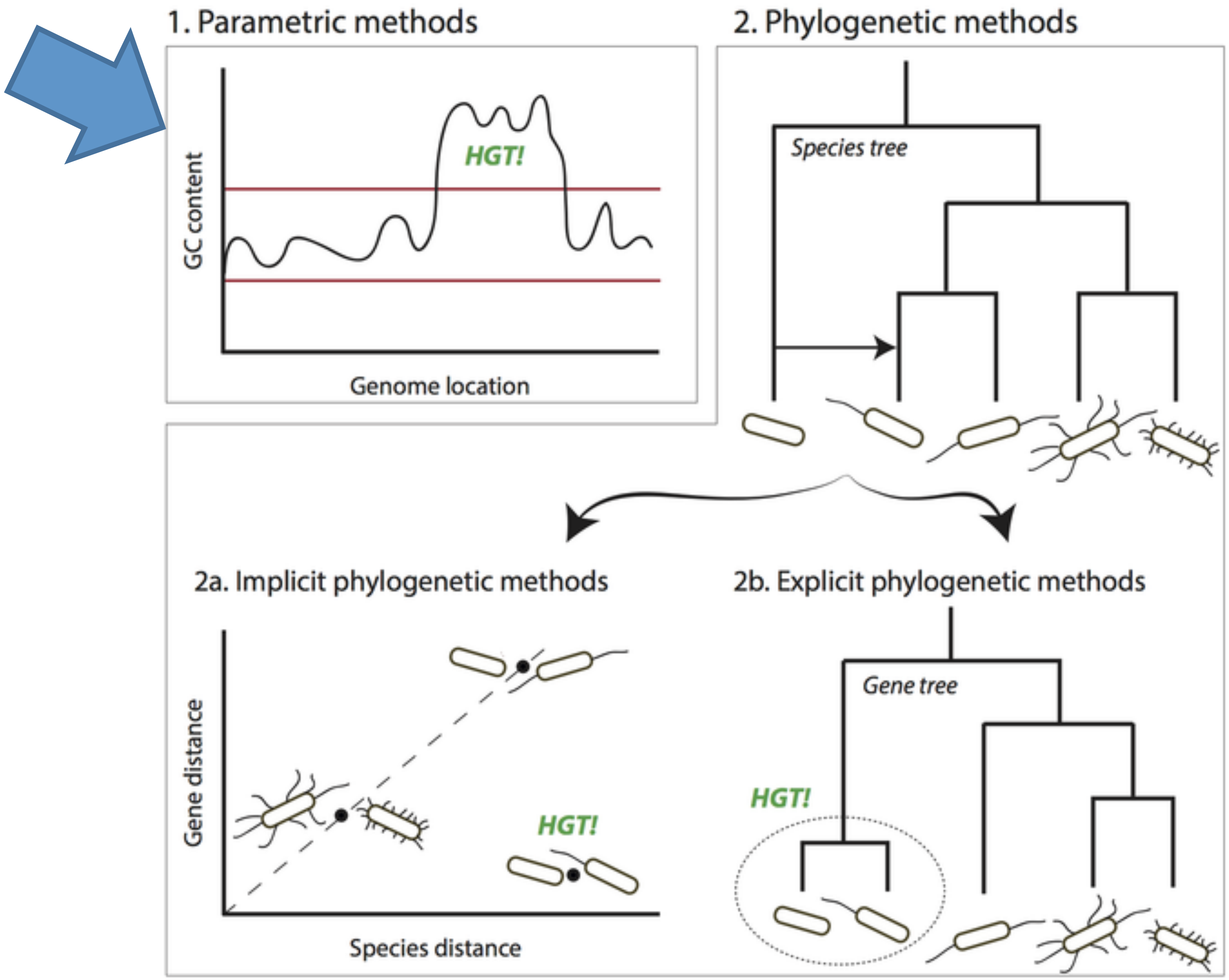
The CLT: central limit theorem



1c The distribution of sampl means

Draw a distribution that can “cheat” CLT

Conceptual overview of HGT inference methods.

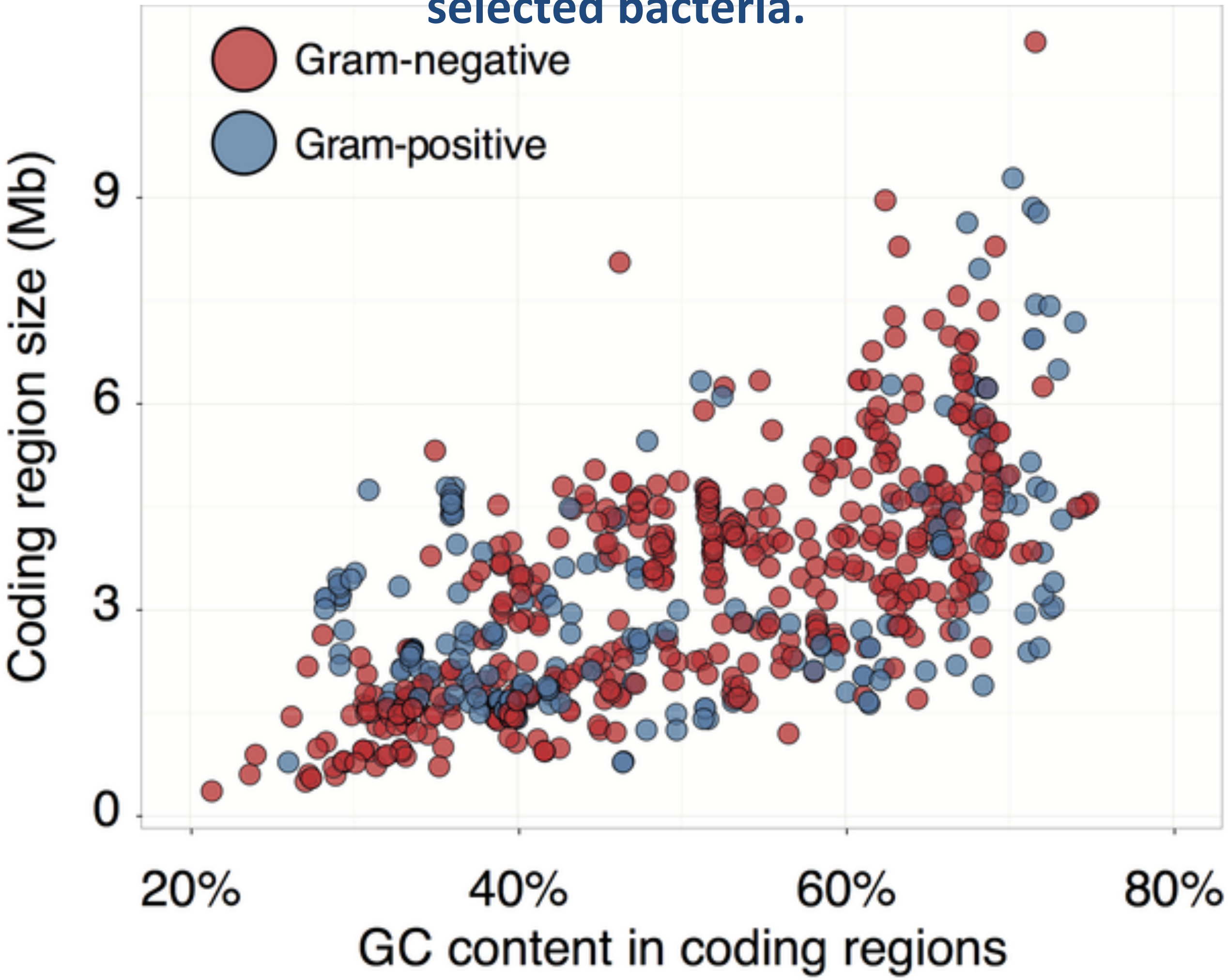


Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015) Inferring Horizontal Gene Transfer. PLOS Computational Biology 11(5): e1004095. https://doi.org/10.1371/journal.pcbi.1004095

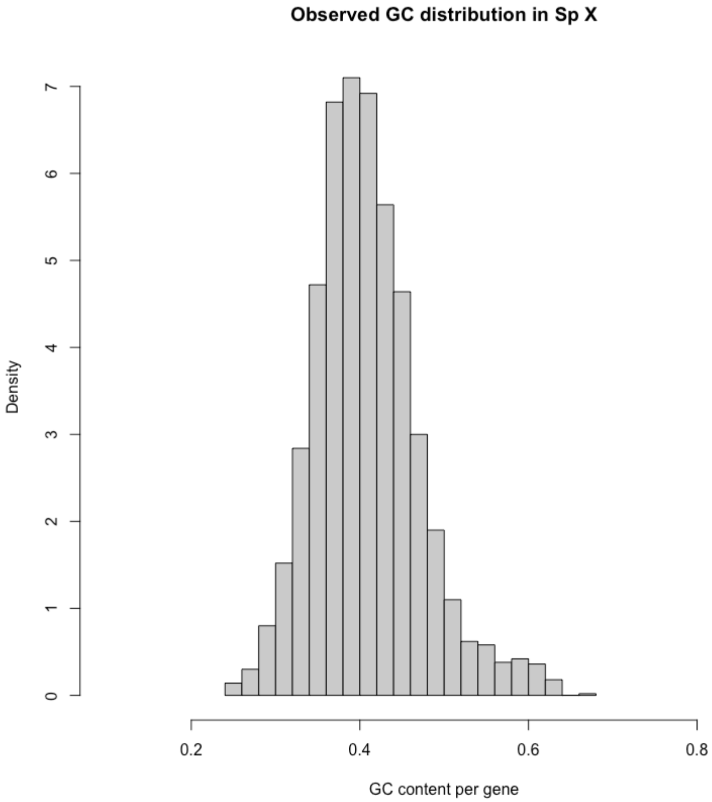
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004095

PLOS | COMPUTATIONAL BIOLOGY

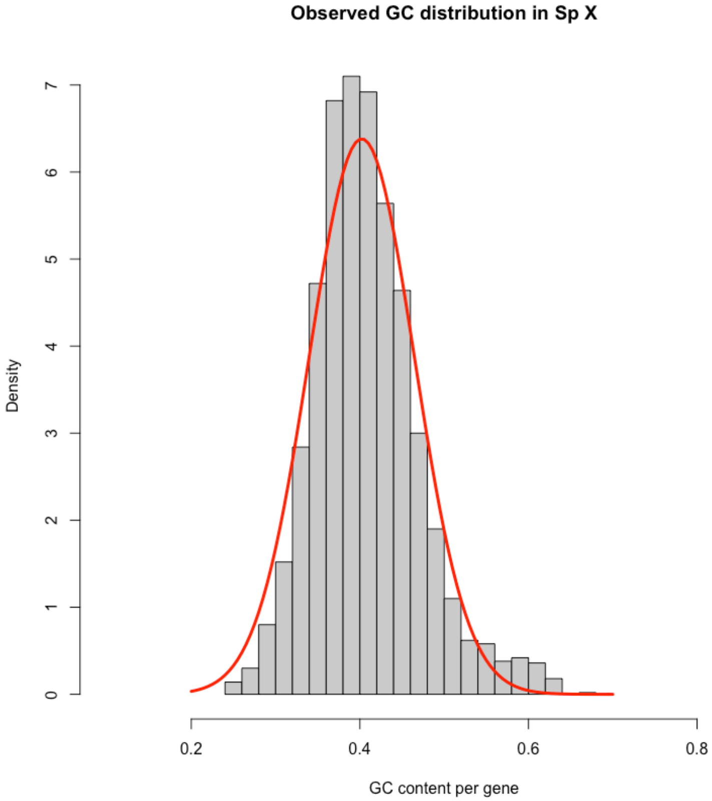
Average GC content of coding regions compared to the genome size for selected bacteria.



Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015) Inferring Horizontal Gene Transfer. PLOS Computational Biology 11(5): e1004095. https://doi.org/10.1371/journal.pcbi.1004095
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004095



1a Properties of the normal distribution



1a Properties of the normal distribution