Bootstrap resampling with independent data

Some theory and applications

Jan Galkowski, Akamai Technologies, inc

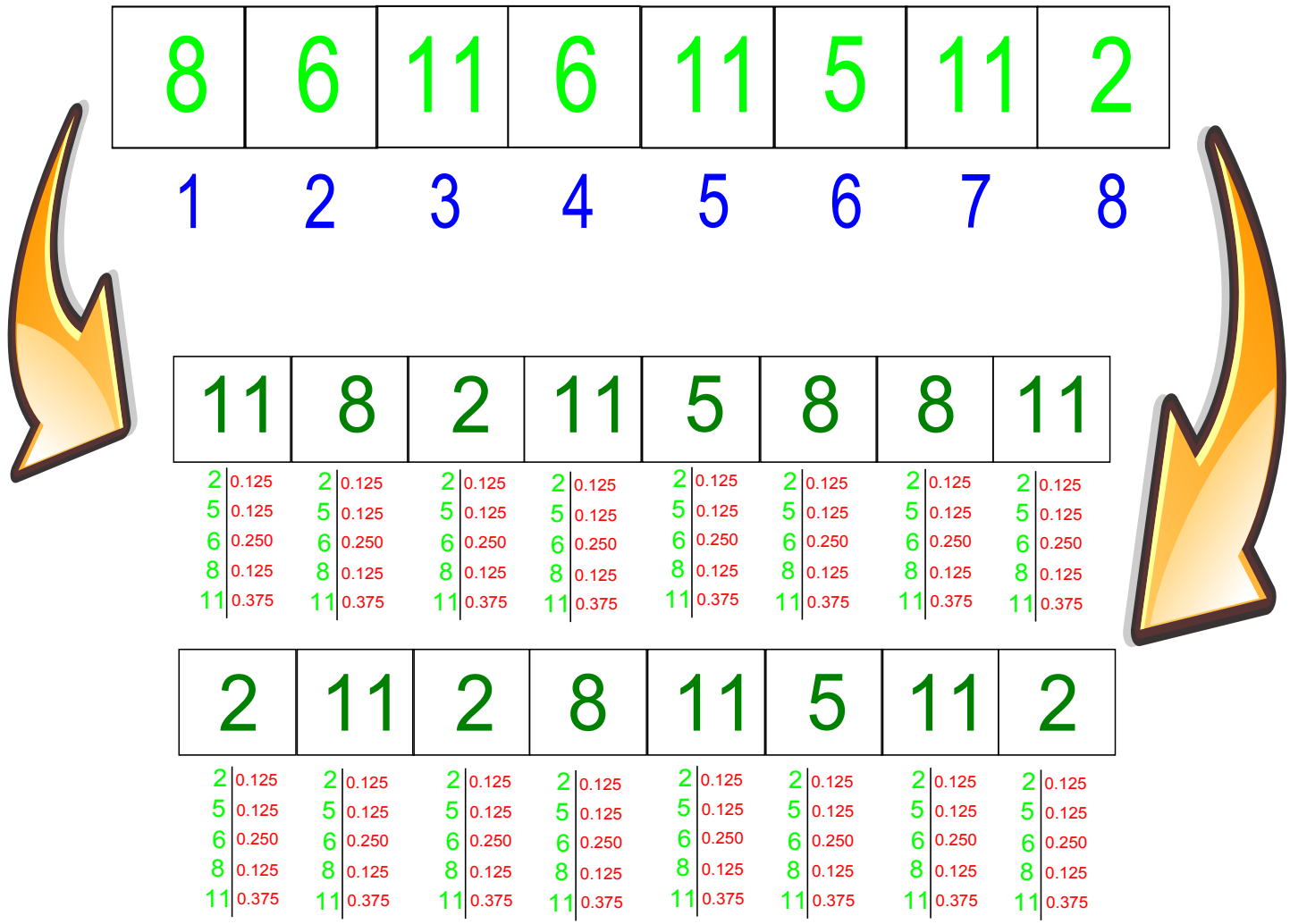
27th November 2012

(last rev November 27, 2012)

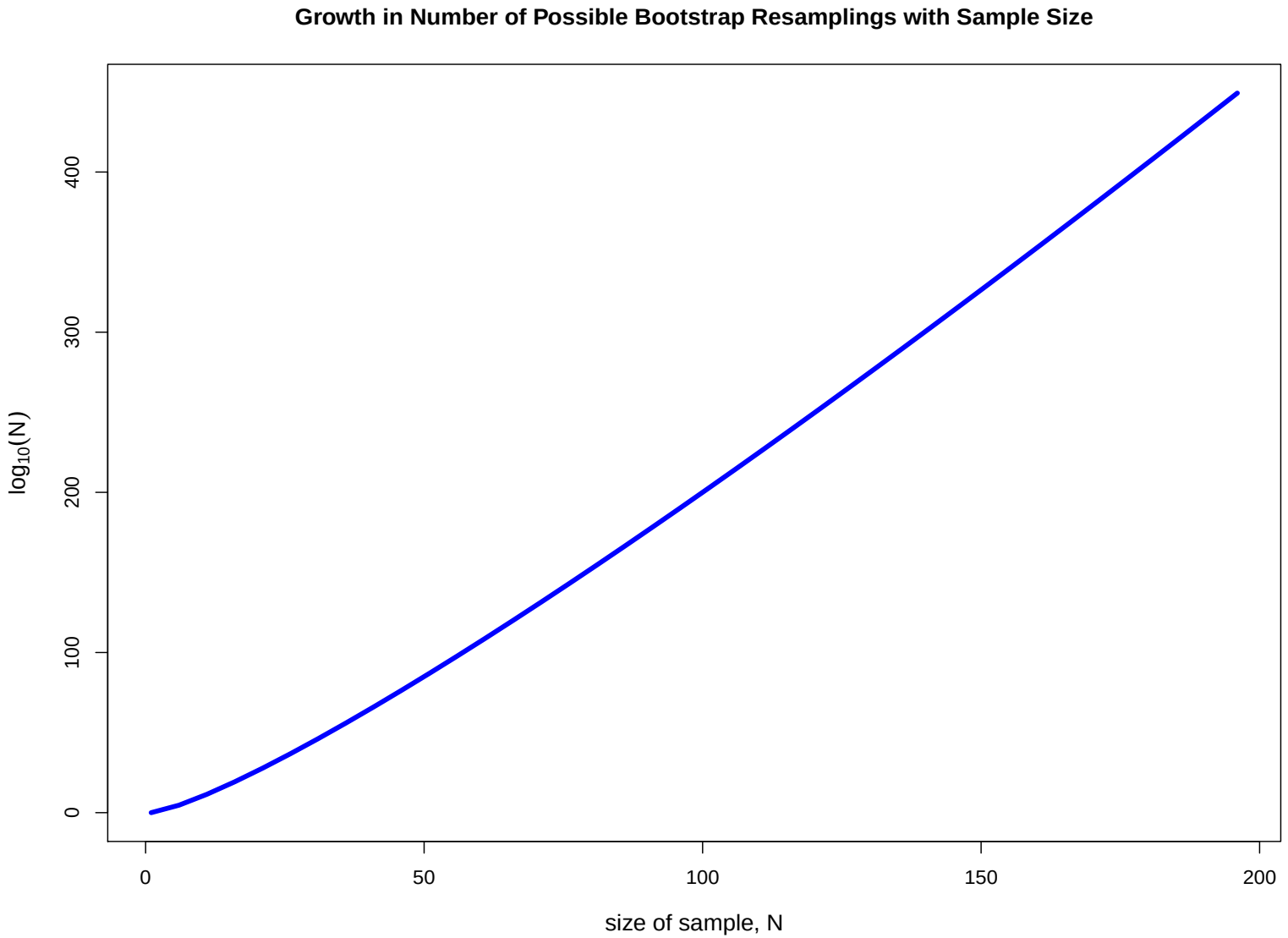
Why do we care about Uncertainty Qualification (UQ)?

- Inherent variability of mechanisms, measurements
- Statistics are themselves random variables, e.g., sample mean, sample standard deviation, sample median
- For simple distributions, knowing the first few moments suffices to characterize them
- For complicated or mixture distributions . . . ?
- We often *pretend* distributions of measurements are stationary. Are they? Diagnostics! What to do if aren't?
- Will have more to say about latter in next lecture on bootstrapping with *dependent data*

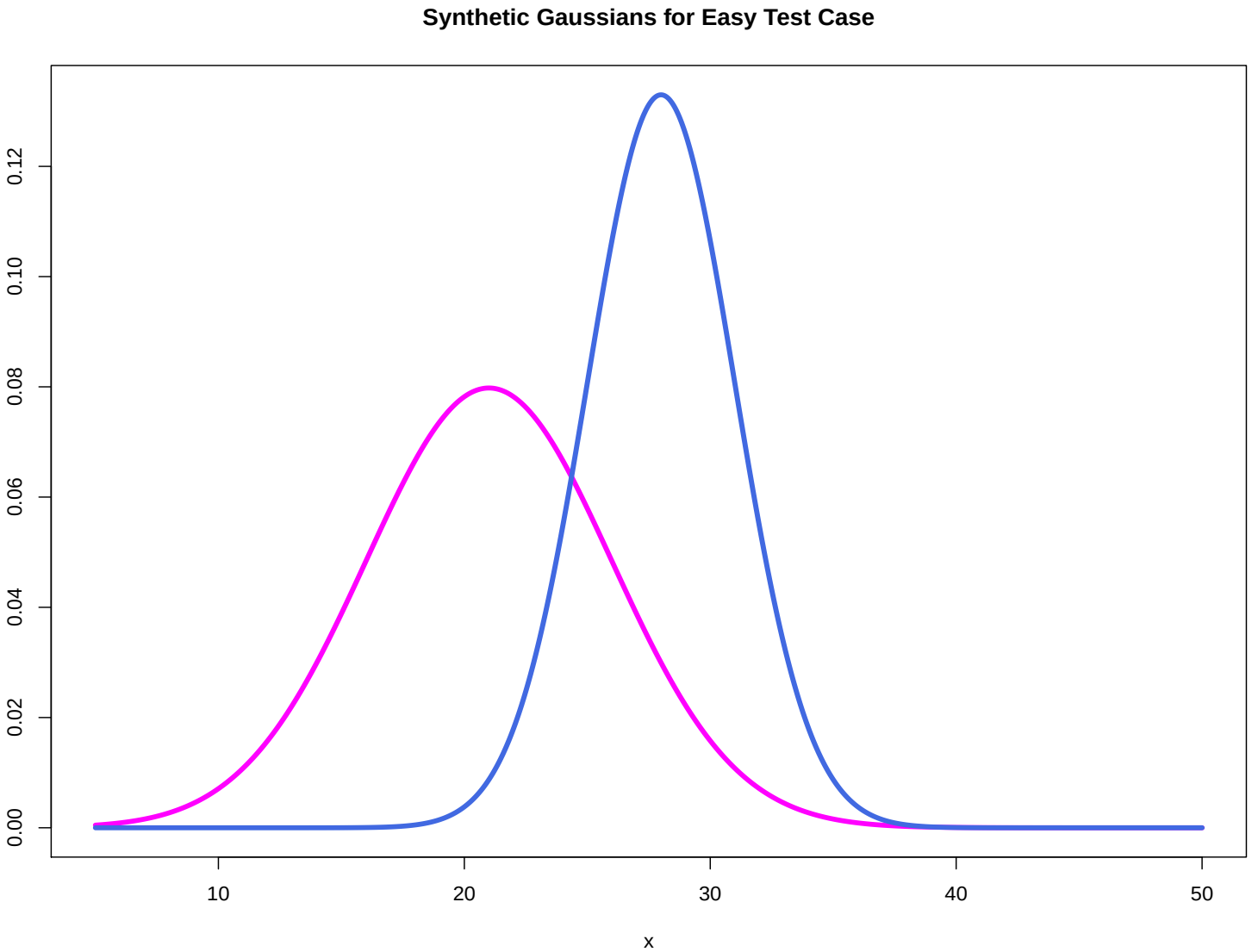
Sketch of How the Bootstrap Works



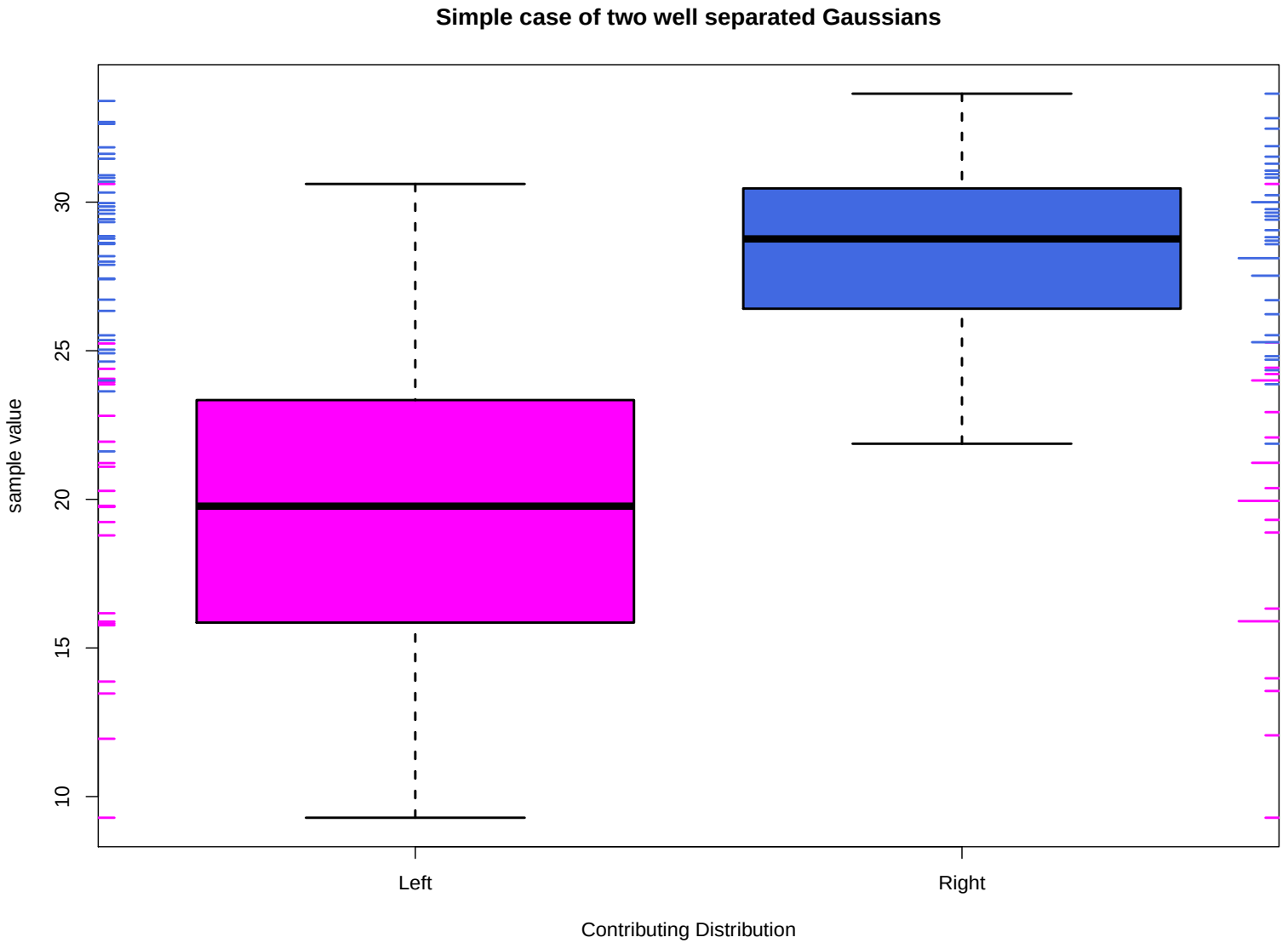
Sketch of Why the Bootstrap Works



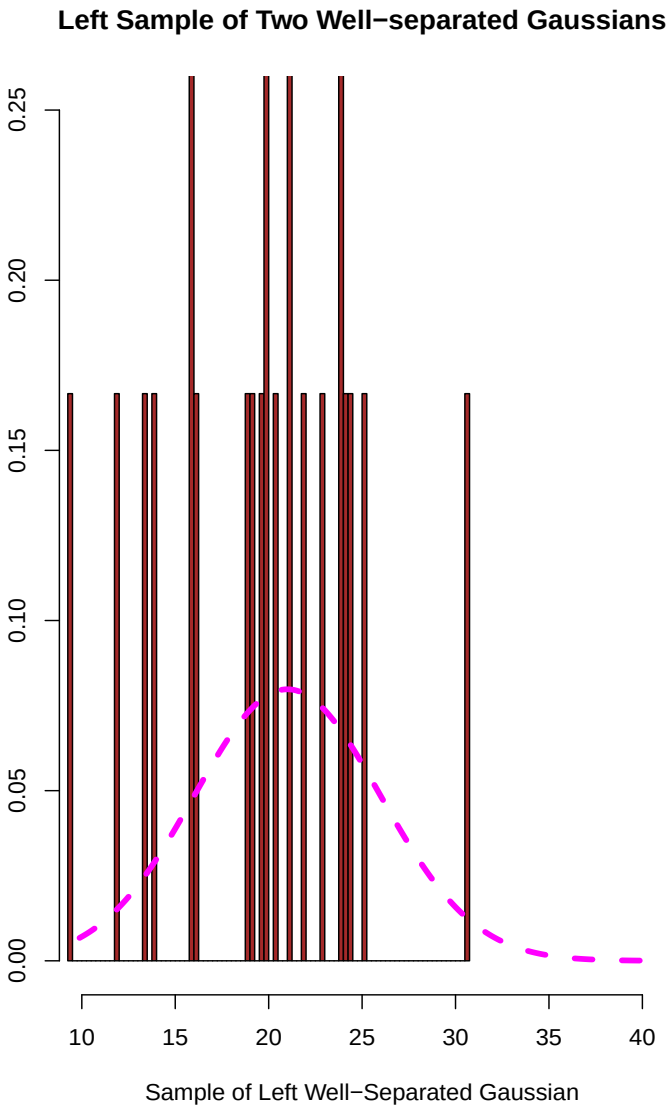
Test Data Model, Easy Case

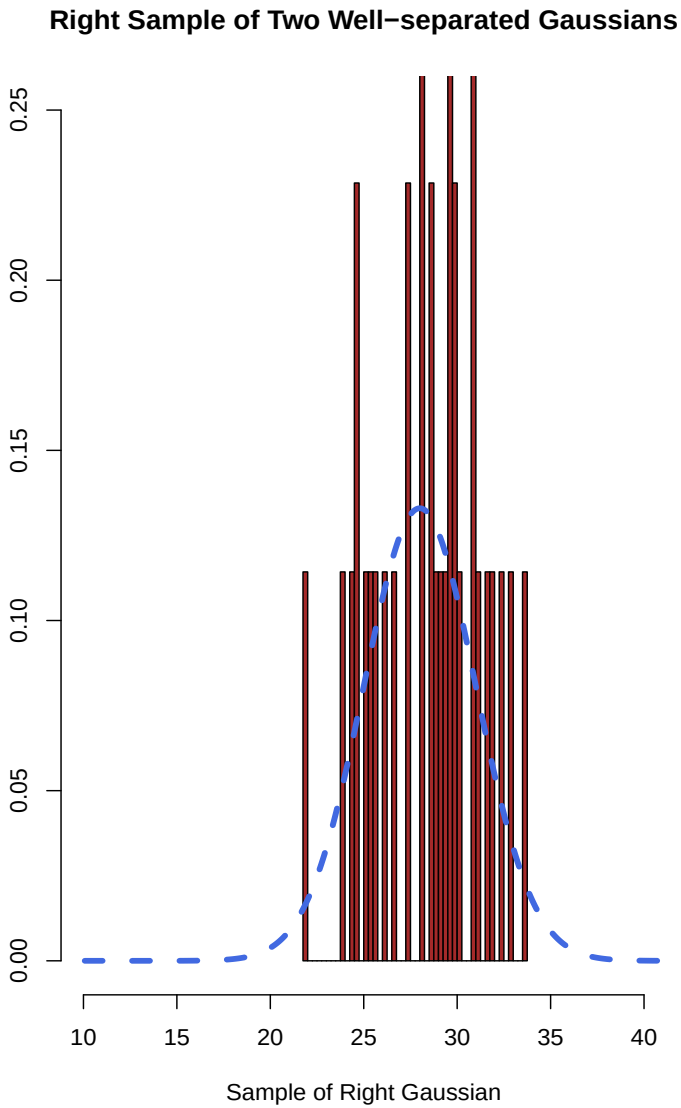


Boxplot of well-separated Gaussians



Histogram of same well-separated Gaussians





Standard *t*-test on well-separated Gaussians

Group	*n*	values average	values s.d.
1: left Gaussians	24	19.540998	4.867407
2: right Gaussians	35	28.471698	2.820691
3: POOLED	59	24.838871	3.782281[1]

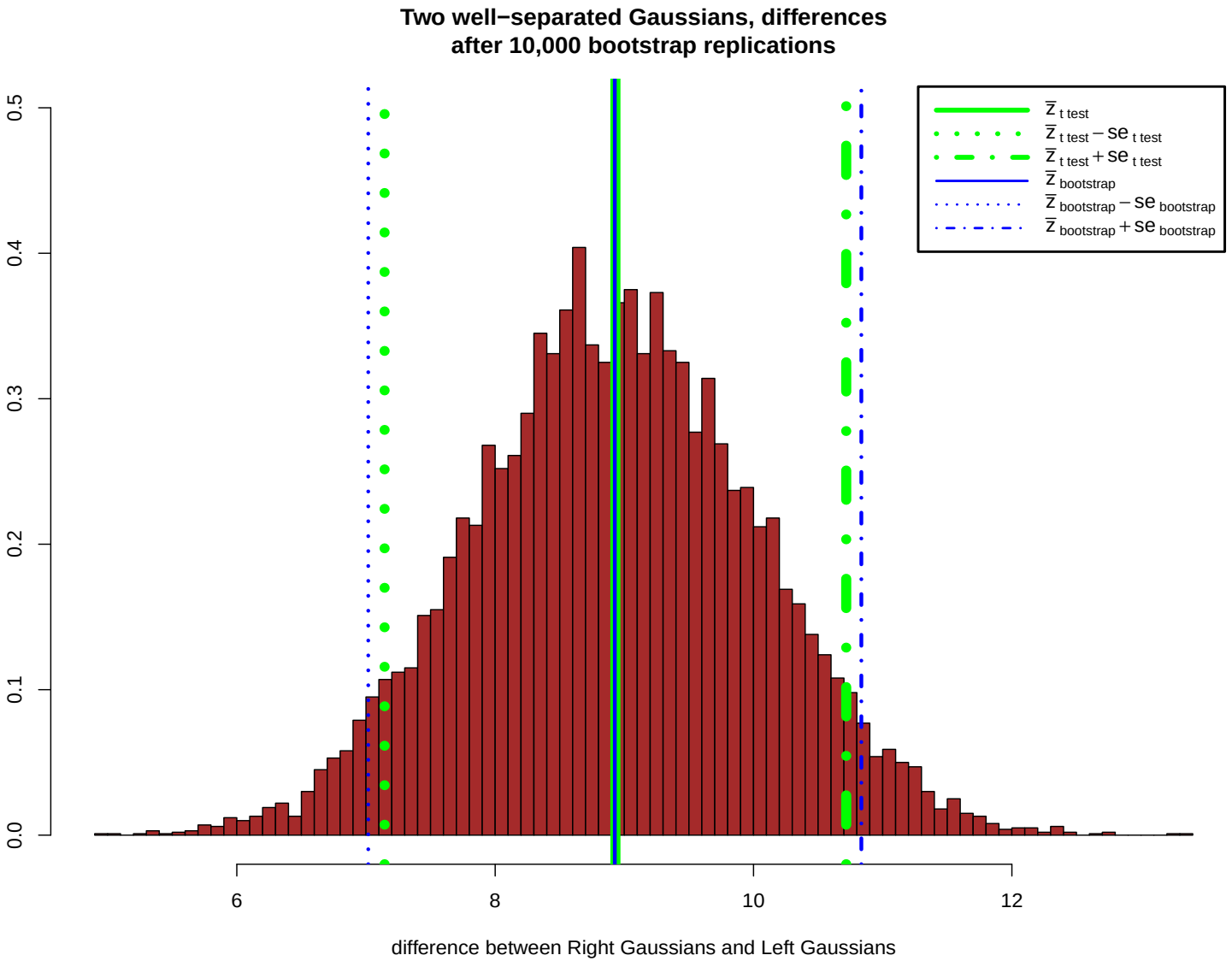
$$(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0} = 8.930700 \qquad \mathrm{SE}(\bar{Y}_2 - \bar{Y}_1) = 1.002398$$

$$\text{two-sided } t\text{-statistic} = \frac{8.930700 - \mathbf{0.0}}{1.002398} = 8.909336$$

$$\mathrm{P} = 0.0000 \because t\text{-statistic} = 8.90934 = t_{57}(0.0000)$$

[1] Adjusted for d.o.f.

Bootstrap characterization of well-separated Gaussians

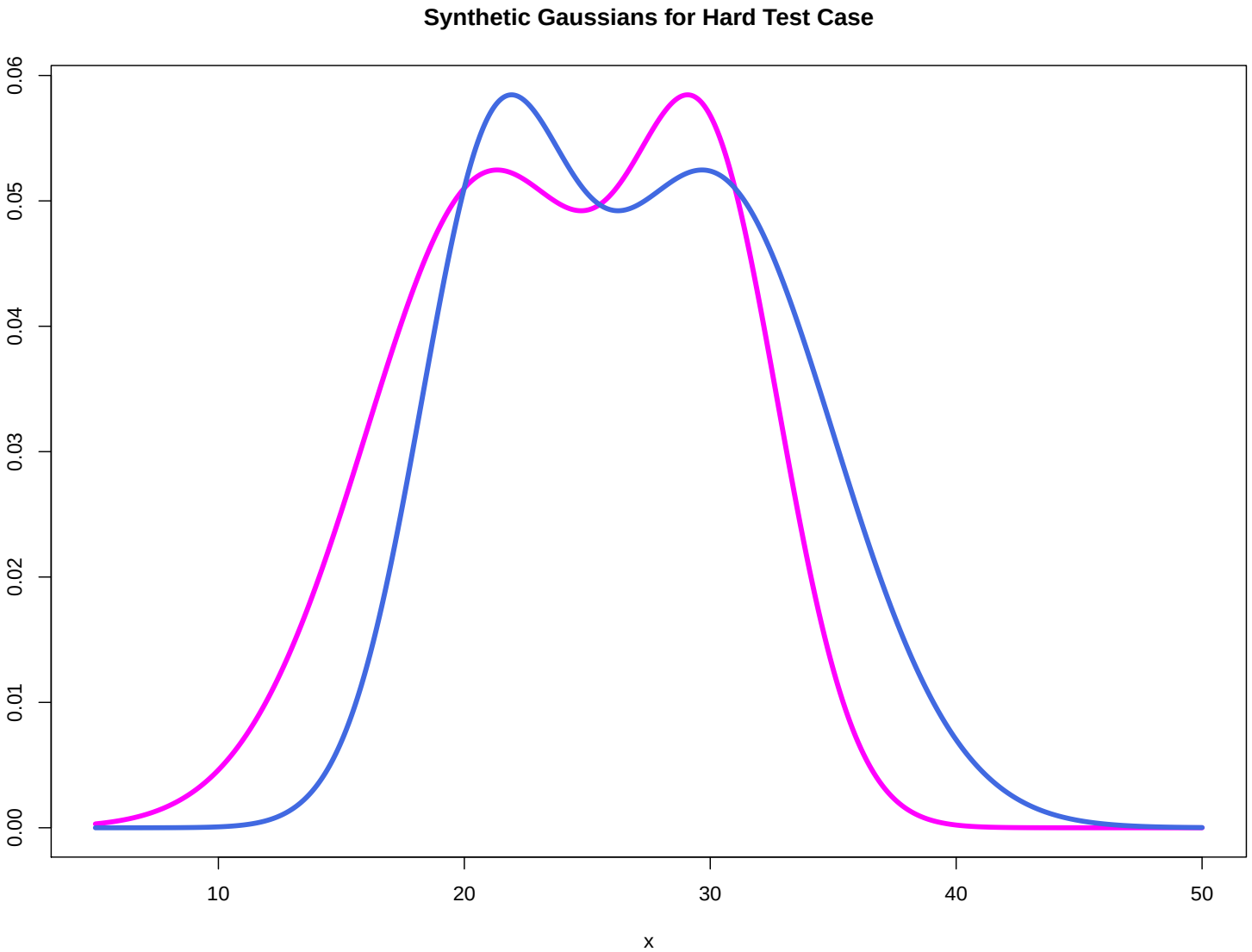


Bootstrap test of well-separated Gaussians

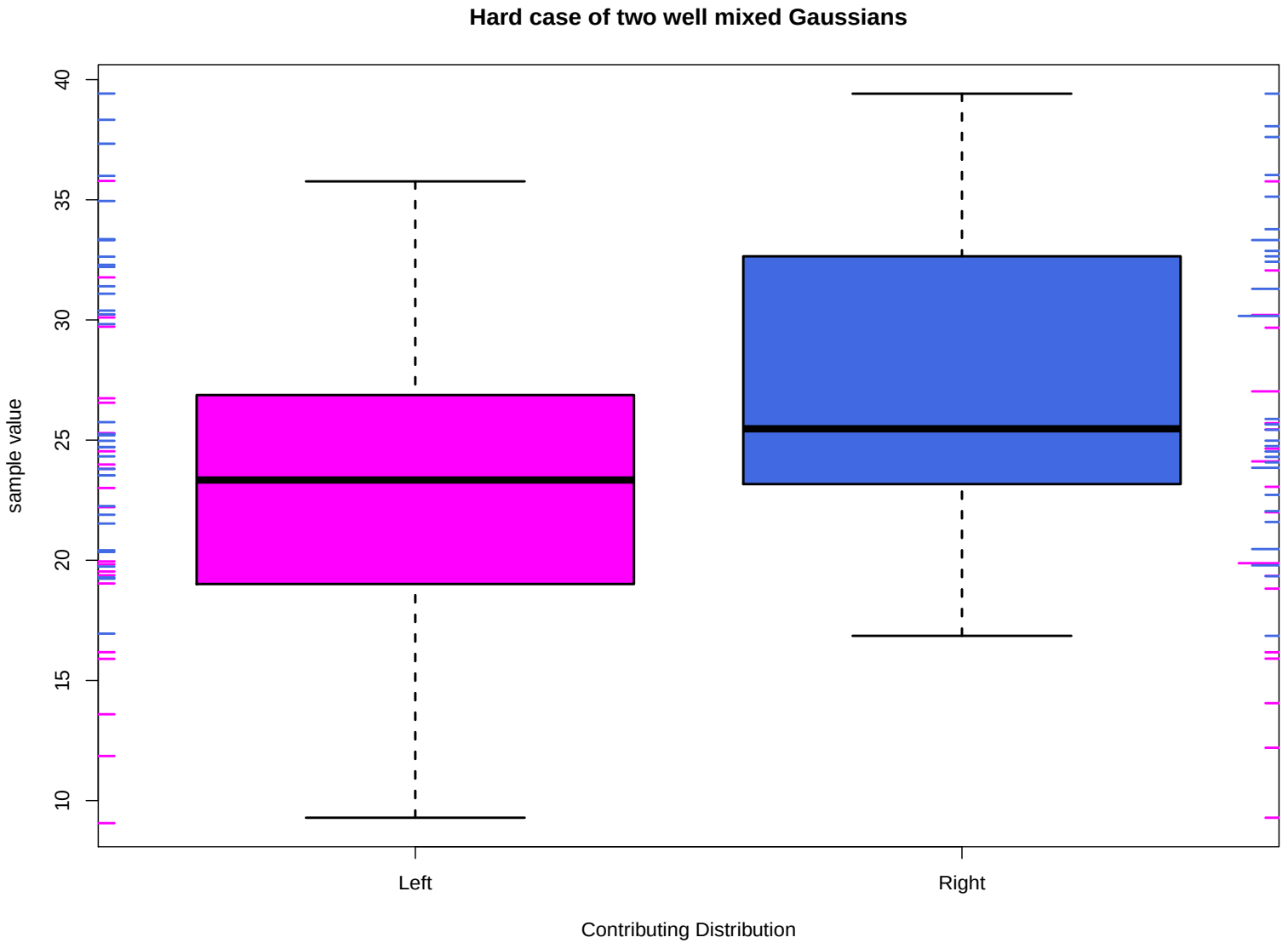
Group	*n*	values average	values s.d.
1: resampled Left	24	19.540998	4.867407
2: resampled Right	35	28.471698	2.820691
3: POOLED	59	24.838871	3.782281[2]
Recall original $(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0}$ = 8.930700			
baseline two-sided t statistic P = 0.0000 $\therefore$ t-statistic = 8.90934 = $t_{57}(0.0000)$			
bootstrap resampled Left two-sided t statistic P = 0.0001 $\therefore$ 0 of 10000 exceed baseline t			
bootstrap resampled Right t statistic P = 0.0002 $\therefore$ 1 of 10000 exceed baseline t			

[2] Adjusted for d.o.f.

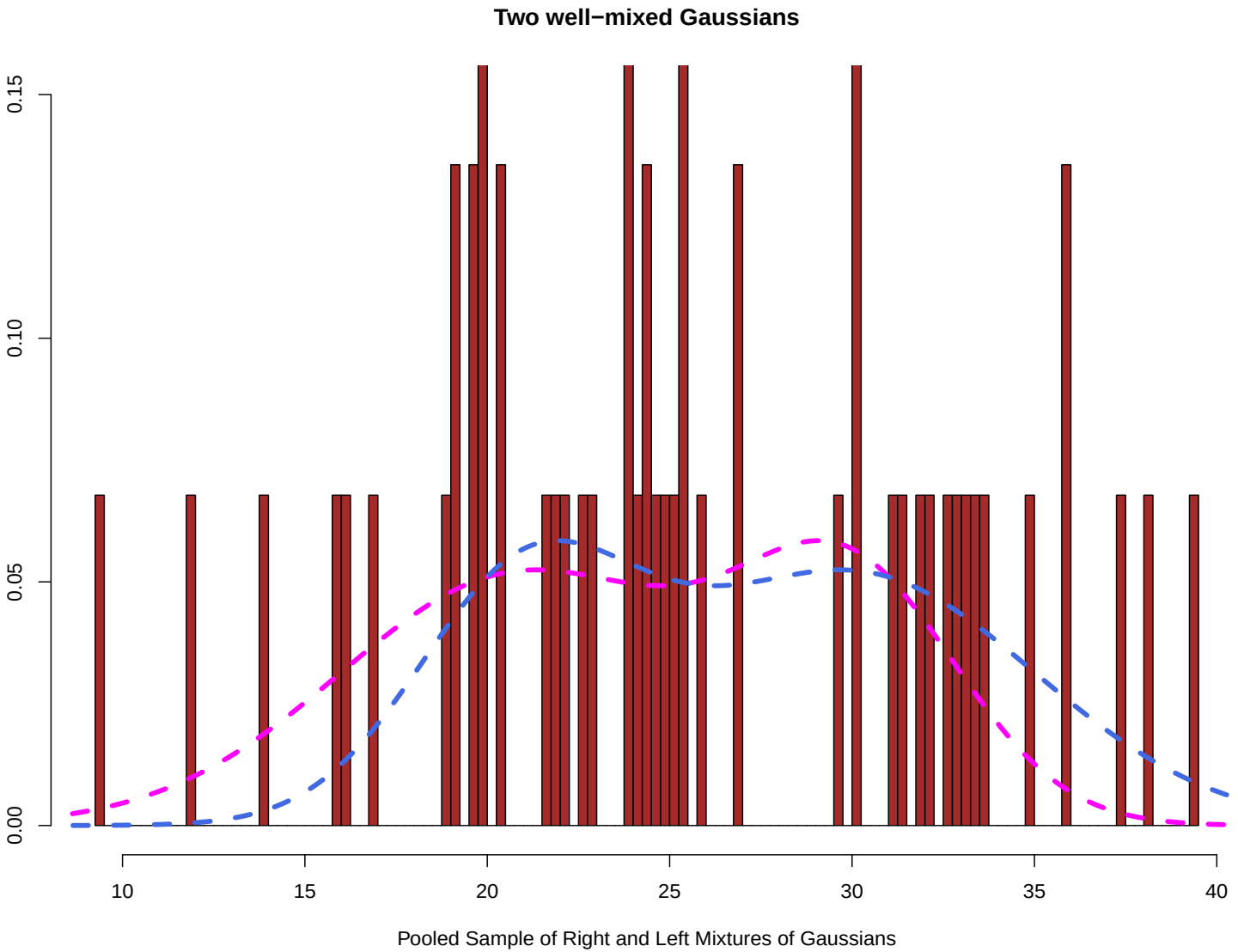
Test Data Model, Hard Case



Boxplot of well-mixed Gaussians



Histogram of same well-mixed Gaussians



Standard *t*-test on mixture distribution of Gaussians

Group	*n*	values average	values s.d.
1: left Gaussians	24	22.635617	6.534154
2: right Gaussians	35	27.558842	6.140611
3: POOLED	59	25.556174	6.302367[3]

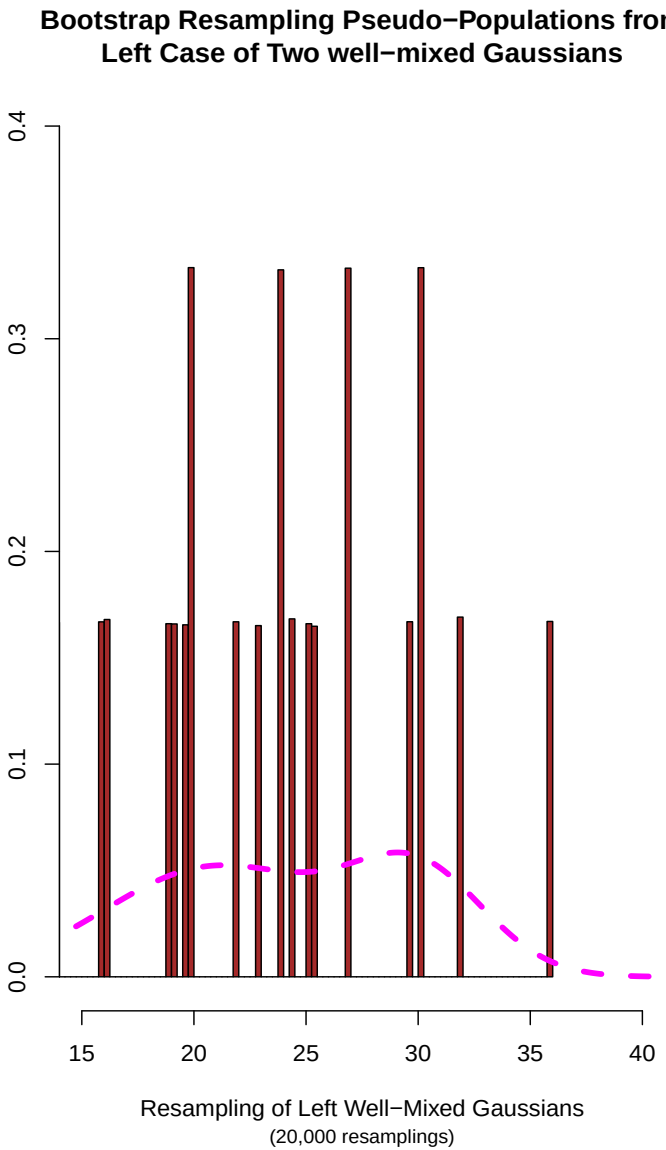
$$(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0} = 4.923225 \qquad \mathrm{SE}(\bar{Y}_2 - \bar{Y}_1) = 1.670283$$

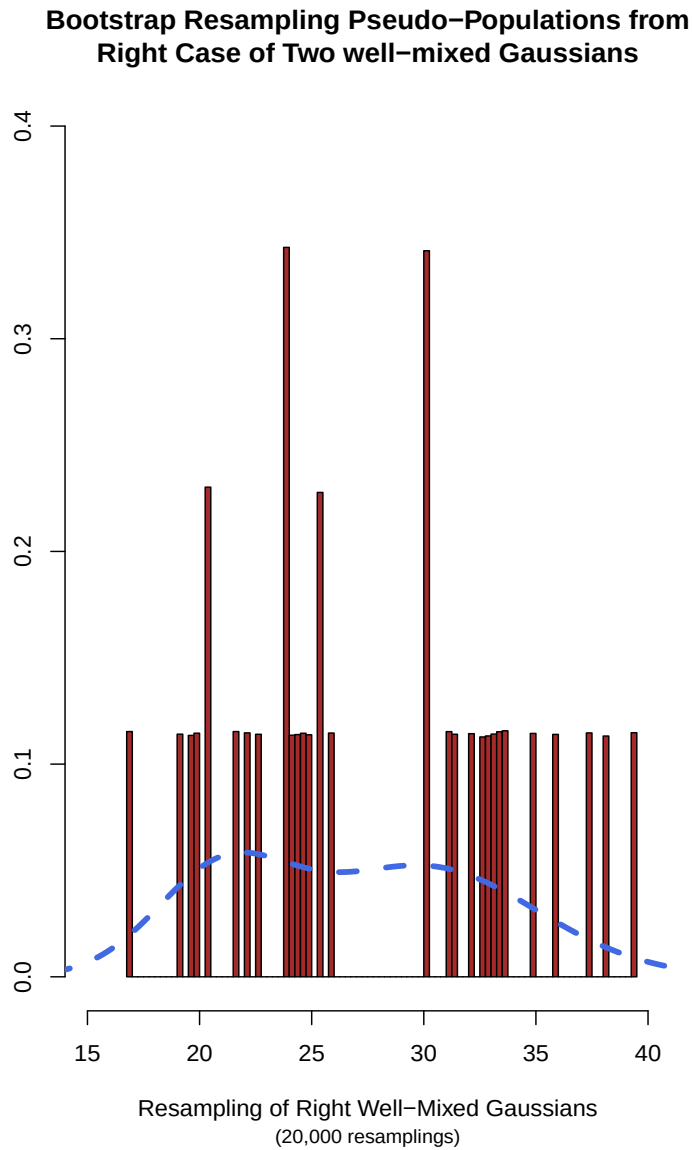
$$\text{two-sided } t\text{-statistic} = \frac{4.923225 - \mathbf{0.0}}{1.670283} = 2.947539$$

$$\mathrm{P} = 0.0023 \because t\text{-statistic} = 2.94754 = t_{57}(0.0023)$$

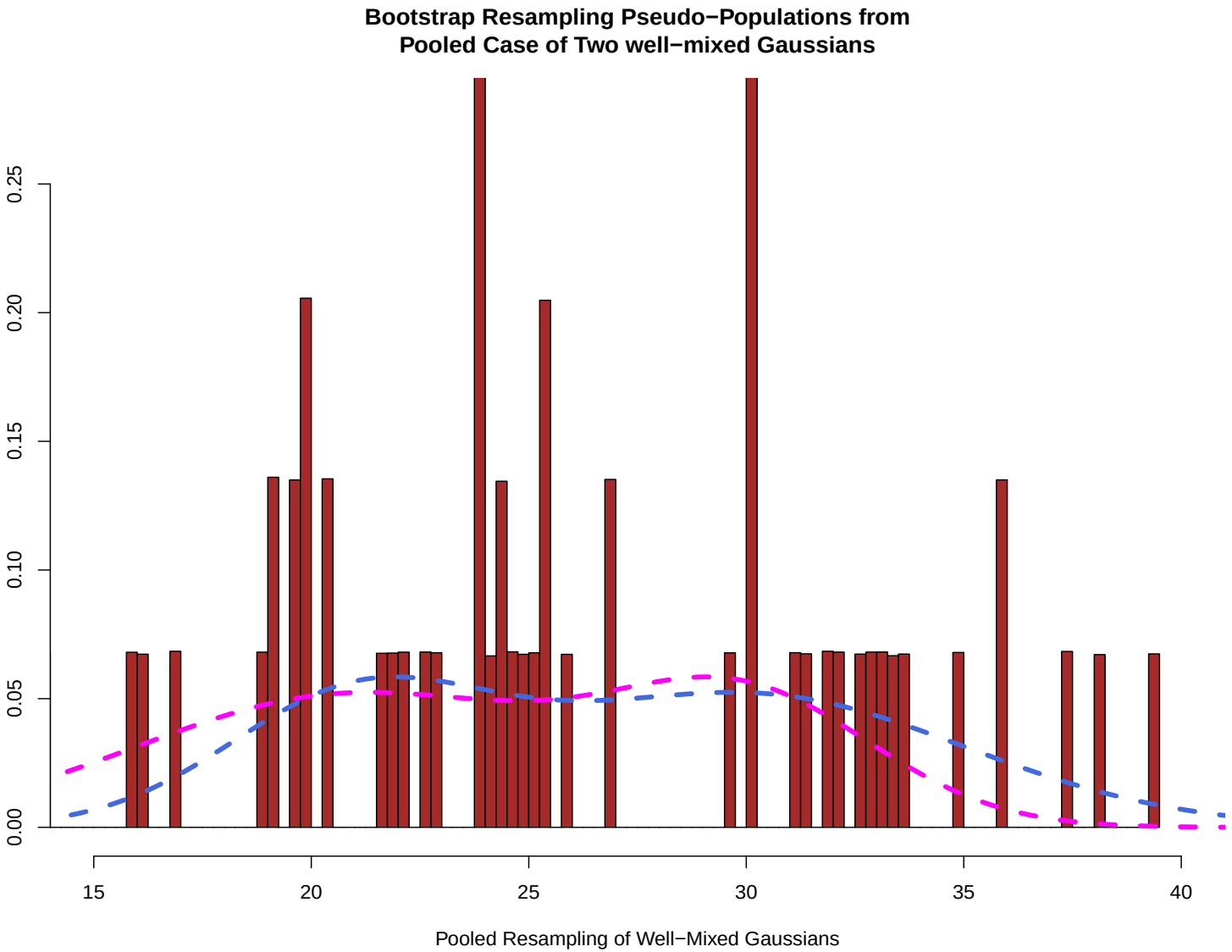
[3] Adjusted for d.o.f.

Resampling from each population separately





What a pooled bootstrap resampling population looks like

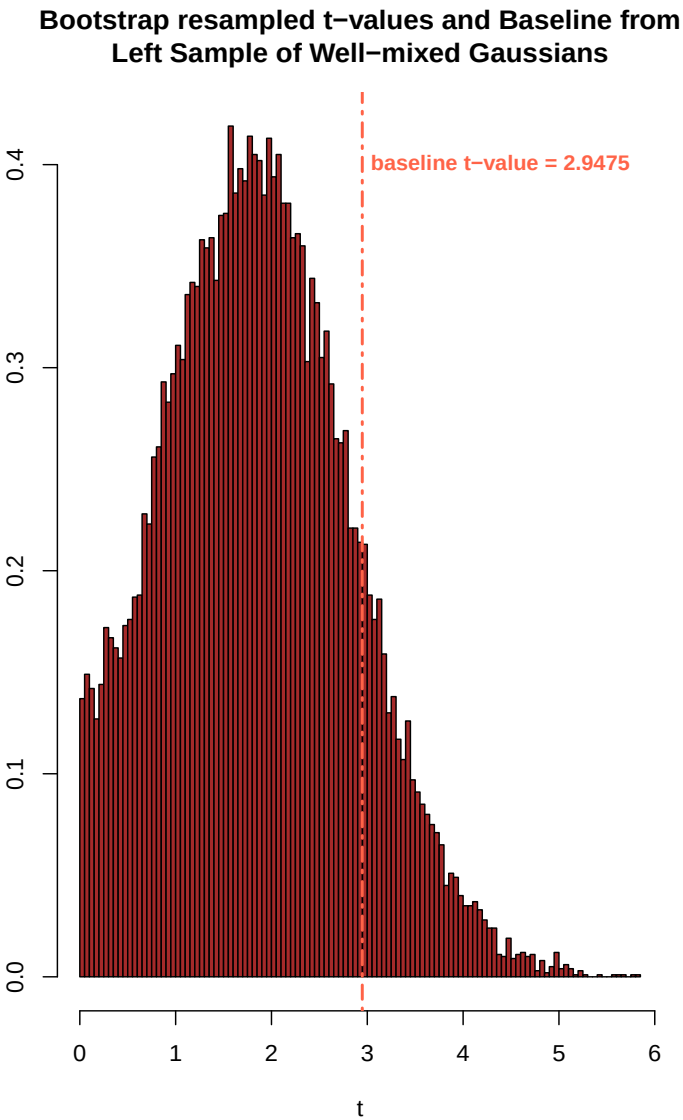


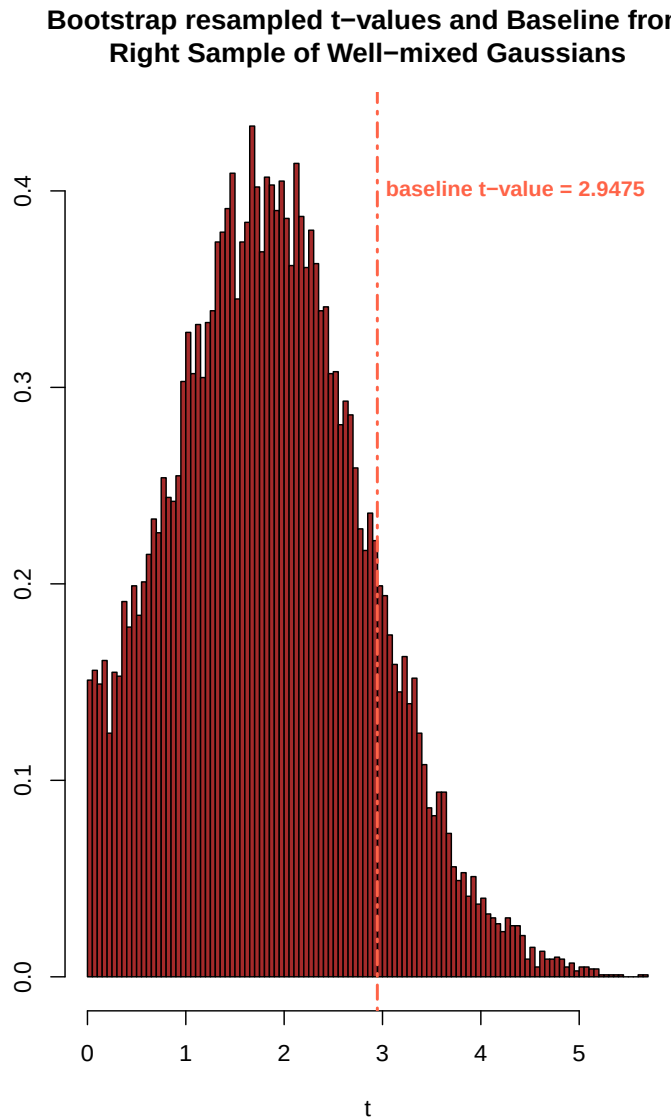
Bootstrap test of well-mixed Gaussians

Group	*n*	values average	values s.d.
1: resampled Left	24	22.635617	6.534154
2: resampled Right	35	27.558842	6.140611
3: POOLED	59	25.556174	6.302367[4]
Recall original $(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0} = 4.923225$			
baseline two-sided t statistic P = 0.0023 $\therefore$ t-statistic = 2.94754 = $t_{57}(0.0023)$			
bootstrap resampled Left two-sided t statistic P = 0.1328 $\therefore$ 2656 of 20000 exceed baseline t			
bootstrap resampled Right t statistic P = 0.1331 $\therefore$ 2662 of 20000 exceed baseline t			

[4] Adjusted for d.o.f.

Resampled t-values





A little Bootstrap theory: Glivenko-Cantelli

- **Empirical distribution function**: i.i.d. $X_1, X_2, \ldots, X_n \sim F$ then

$$F_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}_{X_k \leq x}(x)$$

- **Glivenko-Cantelli Theorem**: As $n \longrightarrow \infty$,

$$\sum_x |F_n - F| \longrightarrow 0$$

almost surely (abbreviated "*a.s.*").

Given

$$\Omega_0 \stackrel{\text{def}}{=} \{\omega : \lim_{n\to\infty} X_n \text{exists}\} = \{\omega : (\limsup_{n\to\infty} X_n) - (\liminf_{n\to\infty} X_n) = 0\}$$

if $P(\Omega_0) = 1$, we say that X_n *converges almost surely*[5].

[5] See Durrett, 4^{th} edition, 2010, pp 15-16.

Proof of Glivenko-Cantelli without jumps in c.d.f.

- Fix x. Let $Y_n = \mathbb{I}_{X_n \leq x}(x)$.
- Since Y_n i.i.d. with $E[Y_n] = P(X_n \leq x) = F(x)$
- Then that and strong laws $\Rightarrow F_n = \frac{1}{n}\sum_{k=1}^{n} Y_k \rightarrow F(x)$ a.s.
- F_n is sequence of nondecreasing functions *that converges pointwise*[6] to *bounded and continuous limit* $F \Rightarrow \sup_u |F_n(u) - F(u)| \rightarrow 0$.

[6] Consider function $\lim_{n\to\infty} cos(\pi x)^{2n}$. This is 1 when x is an integer, and 0 when x is not an integer. Thus, it is discontinuous at every integer and, while it *pointwise convergent*, it is not *uniformly convergent*. Another example: $f_n(x) = x^n$ on $[0,1]$, with $f(x) = 1 \Leftarrow x = 1$ and $f(x) = 0 \Leftarrow 0 \leq x < 1$.

Proof of Glivenko-Cantelli Theorem

After R. Durrett, 2010, 4^{th} edition, *Probability: Theory and Examples*, 76-77

1. Fix x. Let $Z_n = \mathbb{I}_{X_n \leq x}(x)$.
2. Since Z_n i.i.d. with $E[Z_n] = P(X_n < x) = F(x^-) = \lim_{y \to x} F(y)$, strong laws $\Rightarrow F_n(x^-) = \frac{1}{n}\sum_{k=1}^{n} Z_k \to F(x^-)$ a.s.
3. For $1 \leq i \leq k-1$, let $x_{i,k} = \inf\{y : F(y) \geq \frac{i}{k}\}$.
4. Pointwise convergence of $F_n(x)$, $F_n(x^-) \Rightarrow$ can pick W_k so

$$n \geq W_k \Rightarrow |F_n(x_{i,k}) - F(x_{i,k})| < \frac{1}{k}$$

$$n \geq W_k \Rightarrow |F_n(x_{i,k}^-) - F(x_{i,k}^-)| < \frac{1}{k}$$

   for $1 \leq i \leq k-1$.
5. Let $x_{0,k} = -\infty$ and $x_{k,k} = \infty$. Then last inequalities valid for $i = 0$ or $i = k$.
6. Given $u = x_{i-1,k}$ or $u = x_{i,k}$, $1 \leq i \leq k-1$, $n \geq W_k$, monotonicity of F_n, monotonicity of F, $F(x_{i,k}^-) - F(x_{i-1,k}) \leq \frac{1}{k}$,

$$F_n(u) \leq F_n(x_{i,k}^-) \leq F(x_{i,k}^-) + \frac{1}{k} \leq F(x_{i-1,k}) + \frac{2}{k} \leq F(u) + \frac{2}{k}$$

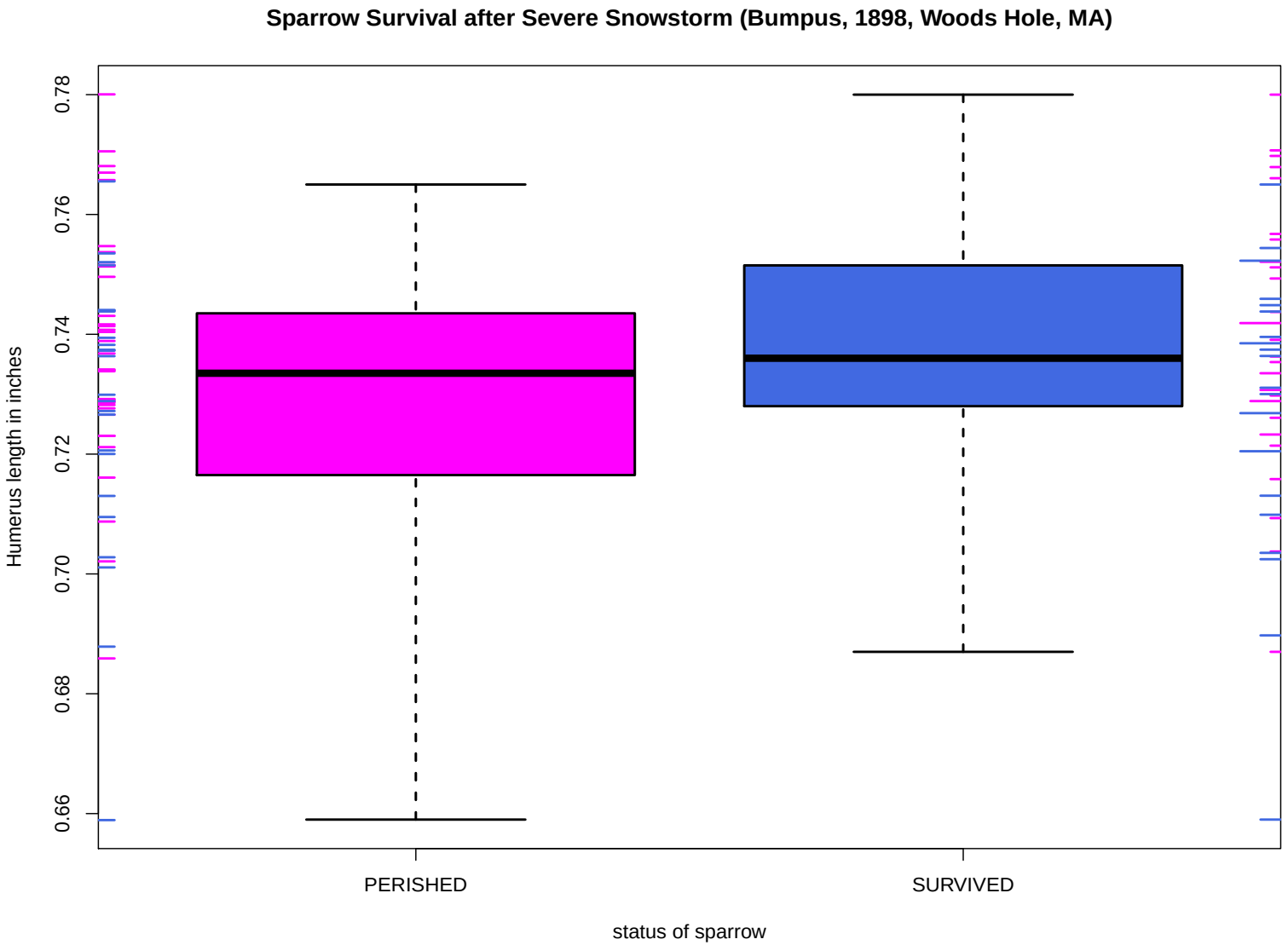
$$F_n(u) \geq F_n(x_{i-1,k}) \geq F(x_{i-1,k}) - \frac{1}{k} \geq F(x_{i,k}^-) - \frac{2}{k} \geq F(u) - \frac{2}{k}$$

7. Consequently, $\sup_u |F_n(u) - F(u)| \leq \frac{2}{k}$. *Result proved.*

A little Bootstrap theory: Edgeworth expansions

- Edgeworth expansion:
  1. $\hat{\theta}$ is constructed from sample of size n
  2. θ_0 is a "parameter"
  3. $\sqrt{n}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, \sigma^2)$
  4. $P\{(\sqrt{n}\frac{\hat{\theta}-\theta_0}{\sigma} \leq x\} = \Phi(x) + (\sqrt{n})^1 p_1(x)\phi(x) + \cdots + (\sqrt{n})^j p_j \phi(x) + \ldots$
  5. $\phi(x) = \mathcal{N}(0,1)$ and $\Phi(x) = \int_{-\infty}^{x} \phi(u)\,\mathrm{d}u$
- Then letting $\hat{A}(\bar{\mathbf{X}}^*) = \frac{g(\bar{\mathbf{X}}^* - g(\bar{\mathbf{X}})}{h(\bar{\mathbf{X}})}$, where s^* denotes the *corresponding bootstrap estimate* of the sample statistic s,
  1. Characteristic function χ of $\mathbf{X}$ satisfies $\limsup_{\|\mathbf{t}\| \to \mathbf{0}} |\chi(\mathbf{t})| < 1$
  2. Asymptotic variance of $\frac{g(\bar{\mathbf{X}}^* - g(\bar{\mathbf{X}})}{h(\bar{\mathbf{X}})} = 1$
  3. Then $P\{\sqrt{n}\frac{g(\bar{\mathbf{X}}^* - g(\bar{\mathbf{X}})}{h(\bar{\mathbf{X}})} \leq x\} = \Phi(x) + \sum_{j=1}^{\nu} (\sqrt{n})^j \pi_j(x)\phi(x) + o(n^{-\nu/2})$
  4. π_j is a polynomial of degree $3j-1$, odd for even j and even for odd j with coefficients a function of moments of $\mathbf{X}$ up to order $j+2$.

Bumpus 1898 data



t-test results from Bumpus 1898 data

Group	***n***	**Humerus average (inches)**	**Humerus s.d. (inches)**
1: perished	24	0.727917	0.023543
2: survived	35	0.738000	0.019839
3: POOLED	59	0.733898	0.021411[7]

$$(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0} = 0.010083 \qquad \mathrm{SE}(\bar{Y}_2 - \bar{Y}_1) = 0.005674$$

$$\text{two-sided } t\text{-statistic} = \frac{0.010083 - \mathbf{0.0}}{0.005674} = 1.776998$$

$$\mathrm{P} = 0.0405 \because t\text{-statistic} = 1.77700 = t_{57}(0.0405)$$

[7] Adjusted for d.o.f.

Bootstrap results from Bumpus 1898 data

Group	***n***	**values average**	**values s.d.**
1: perished	24	0.727917	0.023543
2: survived	35	0.738000	0.019839
3: POOLED	59	0.733898	0.021411[8]
Recall original $(\bar{Y}_2 - \bar{Y}_1) - \mathbf{0.0} = 0.010083$			
baseline two-sided t statistic P = 0.0405 $\therefore$ t-statistic = 1.77700 = $t_{57}(0.0405)$			
bootstrap perished two-sided t statistic P = 0.2633 $\therefore$ 2632 of 10000 exceed baseline t			
bootstrap survived t statistic P = 0.2633 $\therefore$ 2632 of 10000 exceed baseline t			

[8] Adjusted for d.o.f.

Can go elsewhere: Stroke risk and proportions

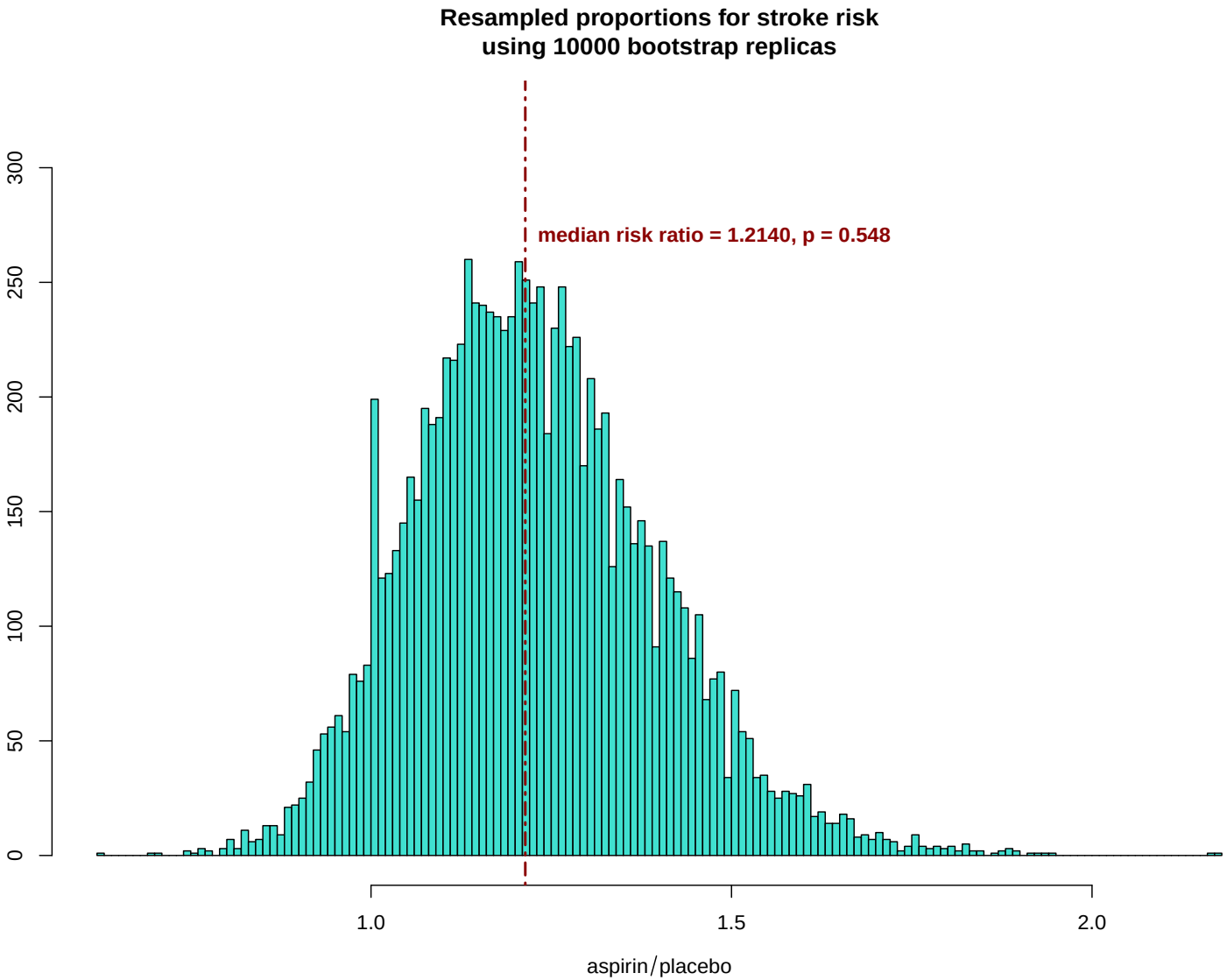
	strokes[9]	subjects
aspirin group	119	11,037
placebo group	98	11,034

How to bootstrap this problem . . .

1. Create an “aspirin sample” of length 11,037 having 119 ones and 11,037 - 119 zeros.
2. Create a “placebo sample” of length 11,034 having 98 ones and 11,037 - 98 zeros.
3. Draw *with replacement* of length 11,037 from “aspirin sample”.
4. Draw *with replacement* of length 11,034 from “placebo sample”.
5. Calculate proportion ones in draw from “aspirin sample” to number of ones in draw from “placebo sample”.
6. Repeat drawing and calculation a large number of times, say, 1000.
7. Calculate the sample statistics you like from the resulting population.

[9] From *New York Times* of 27th January 1987, as quoted by Efron and Tibshirani in *An Introduction to the Bootstrap*, 1998.

Bootstrap display of proportions in stroke risk



Can go elsewhere: In lieu of cross-validation

- *Cross-validation* is a bag of techniques for trying to assess *prediction error*.
- There are other "bags of techniques" as well, one will be mentioned later, but a review needs await a lecture on *predictive inference*.

$$\mathbf{y} + \delta_{\mathbf{y}} = [\mathbf{f} + \delta_{\mathbf{f}}](\mathbf{x} + \delta_{\mathbf{x}}) \longrightarrow \delta_{\mathbf{y}} = \mathbf{f}(\delta_{\mathbf{x}}) + \delta_{\mathbf{f}}(\mathbf{x} + \delta_{\mathbf{x}})$$

- This is gotten wrong a lot. Important to separate variation in *data* from variation in model.
- Will present the standard set of methods, and then offer one Bootstrap-based alternative.

Cross-validation: A review

Classical methods

- **Split-half**:
  1. *Randomly* pick half data subset from sample without replacement.
  2. Fit to that.
  3. Test fit on half *not* selected.
  4. Report measure like MSE for test.
- K**-fold**: Let $F = \{k : k = 1, \ldots, K\}$.
  1. Divide sample into K *randomly picked* disjoint subsets.
  2. For each $j, j \in F$, fit to all $i \neq j, i \in F$.
  3. Then test the fit obtained on subset j.
  4. Report measure like MSE by averaging over all K of the "folds".
- **Leave one out cross-validation**: K-fold with $K = 1$. (Not covered here.)

Cross-validation: A review

Bootstrap-related methods

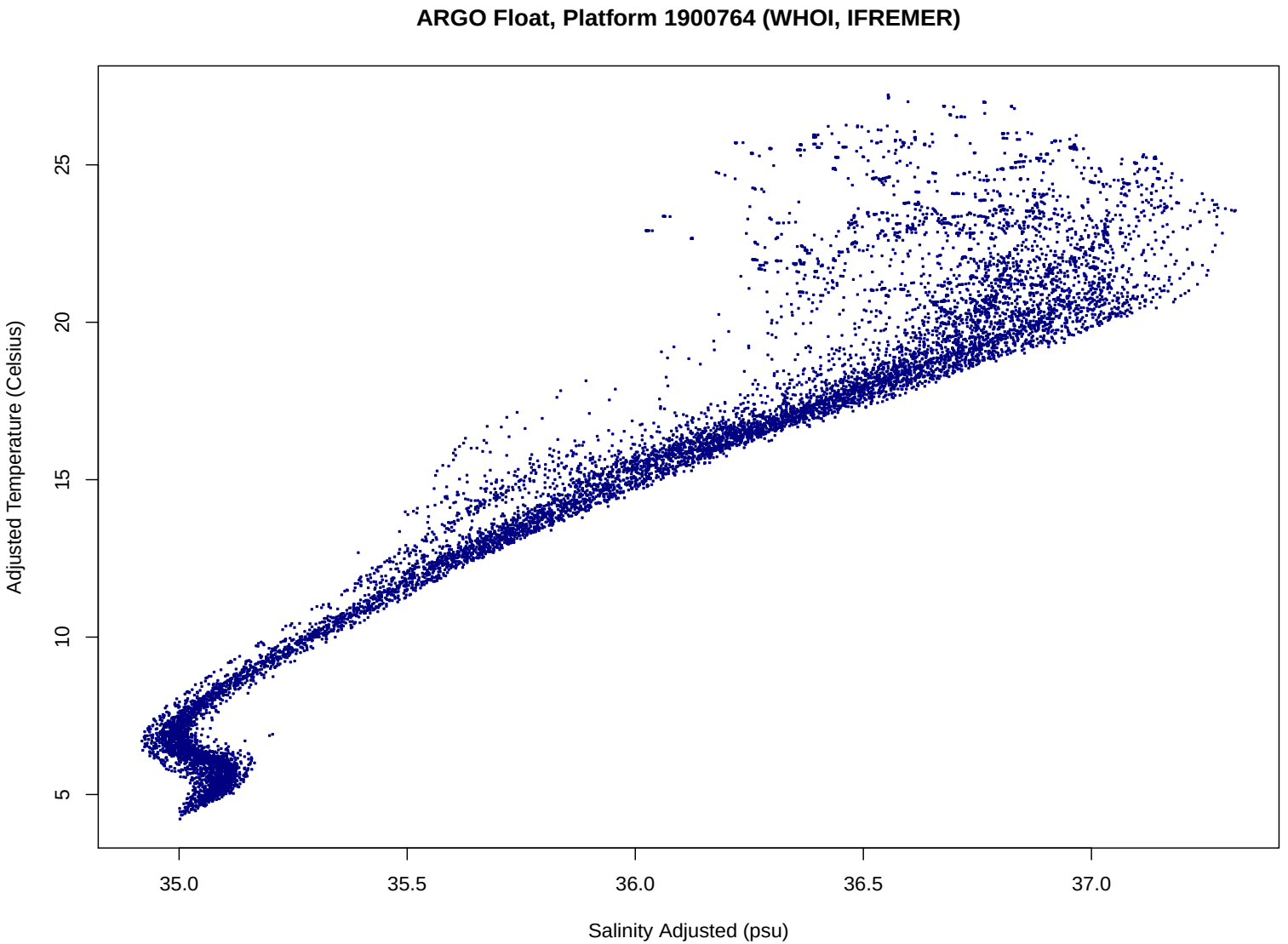
- **Bootstrap validation**:
  1. Take *B* Bootstrap resamplings from the sample.
  2. For each resampling, fit the model to the resampling, then predict the model on the resampling *and on the entire original sample*.
  3. Note the difference between performance on the entire original sample and the performance on the resampling in each case. Each of these values are called an estimate of *optimism*.
  4. Average the optimism values, and then add that to the average residual squared error of the original fit to get an estimate of prediction error.

Two versions of Bootstrap-based cross validation illustrated here. The first is the "ordinary Bootstrap" realization of validation[10], and the "0.632+ bootstrap", a refined version designed to replace classical cross-validation.

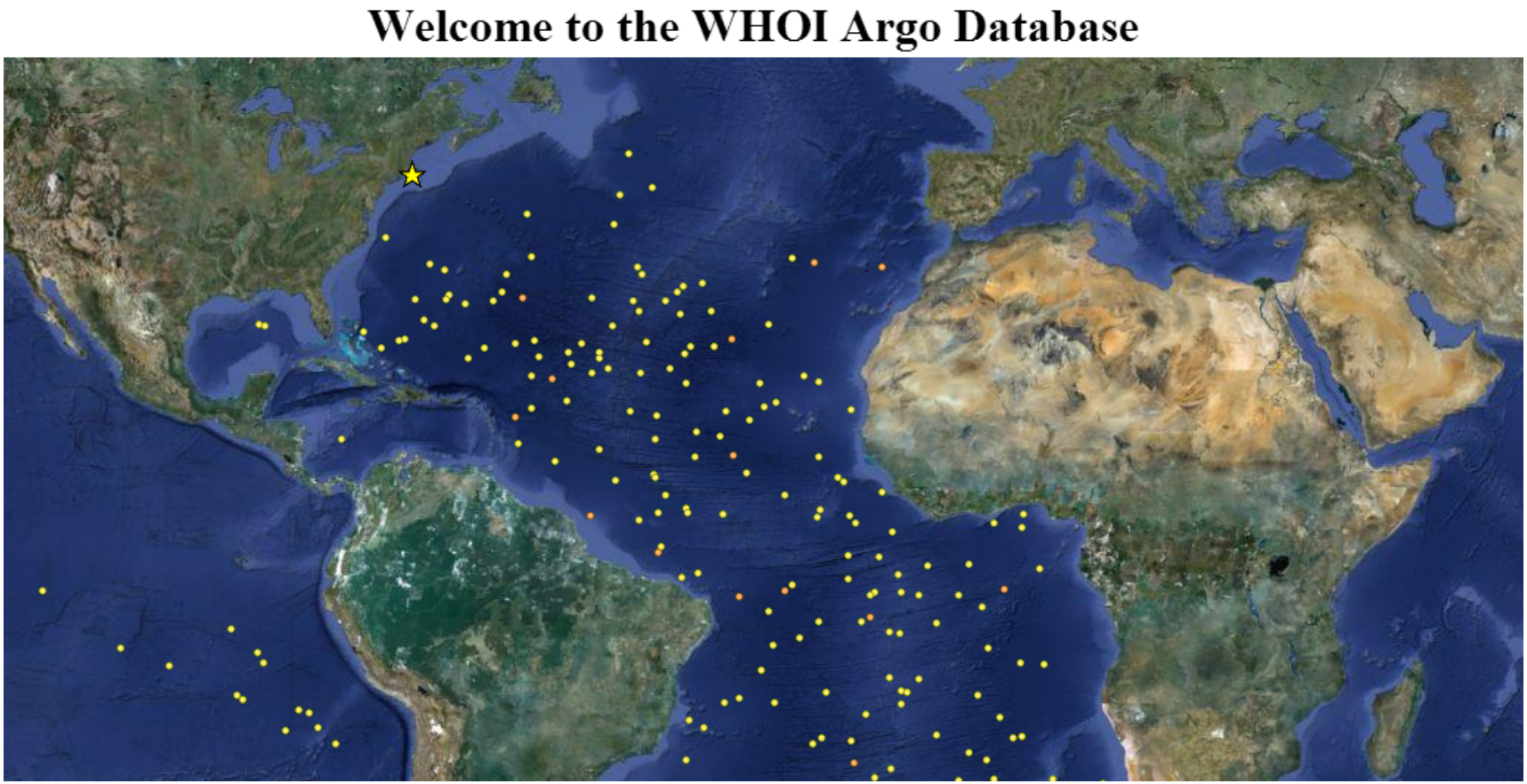
[10] B. Efron, G. Gong, "A leisurely look at the Bootstrap, the Jackknife, and cross-validation", *The American Statistician*, 37(1), 1983, 36-48, as realized by F. E. Harrell, Jr, in the **R** package *rms*.

Data to model: Salinity as predicted by temperature, pressure, latitude, month

ARGO float, platform 1900764 (WHOI, IFREMER)

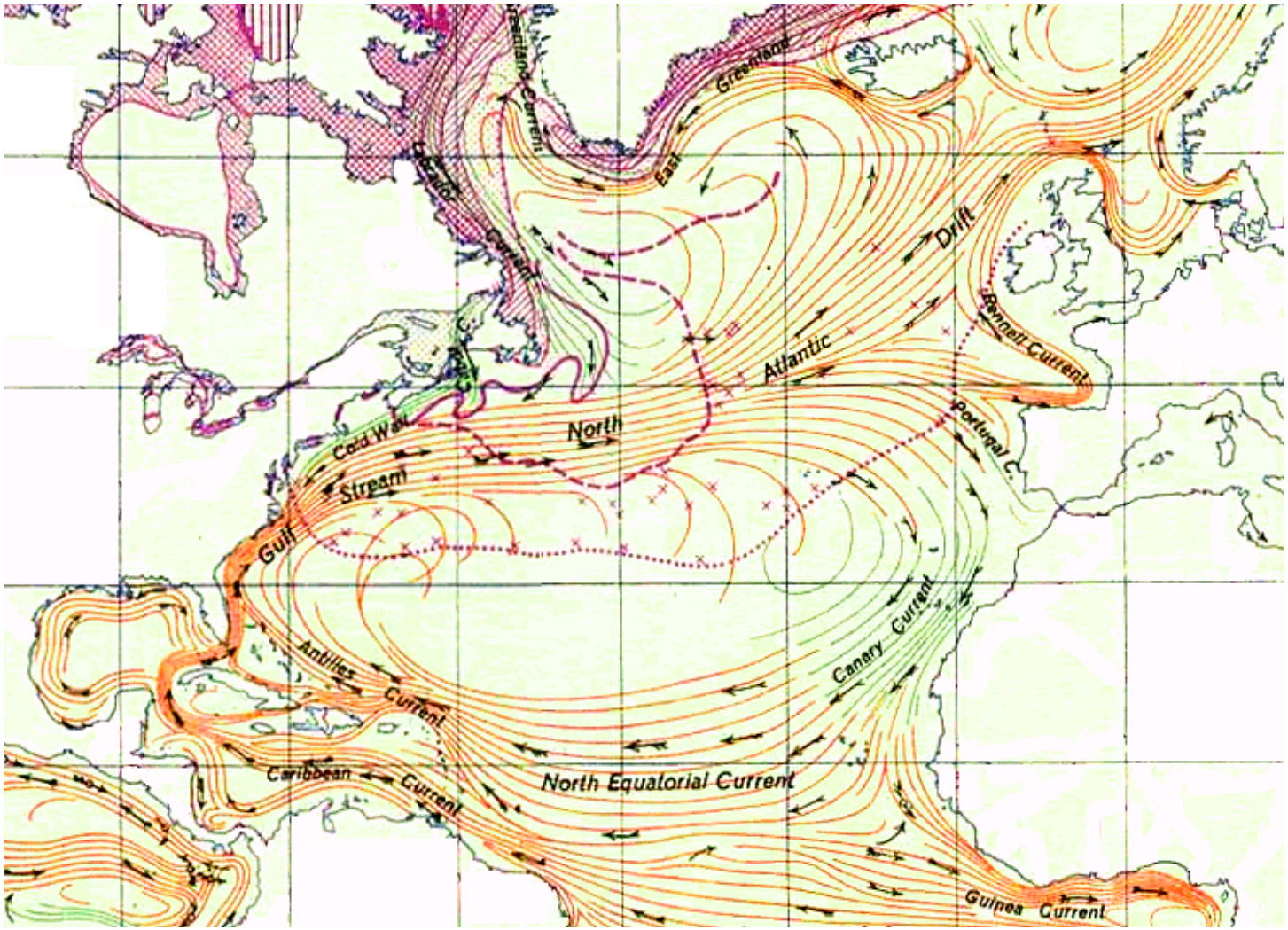


The ARGO Floats System

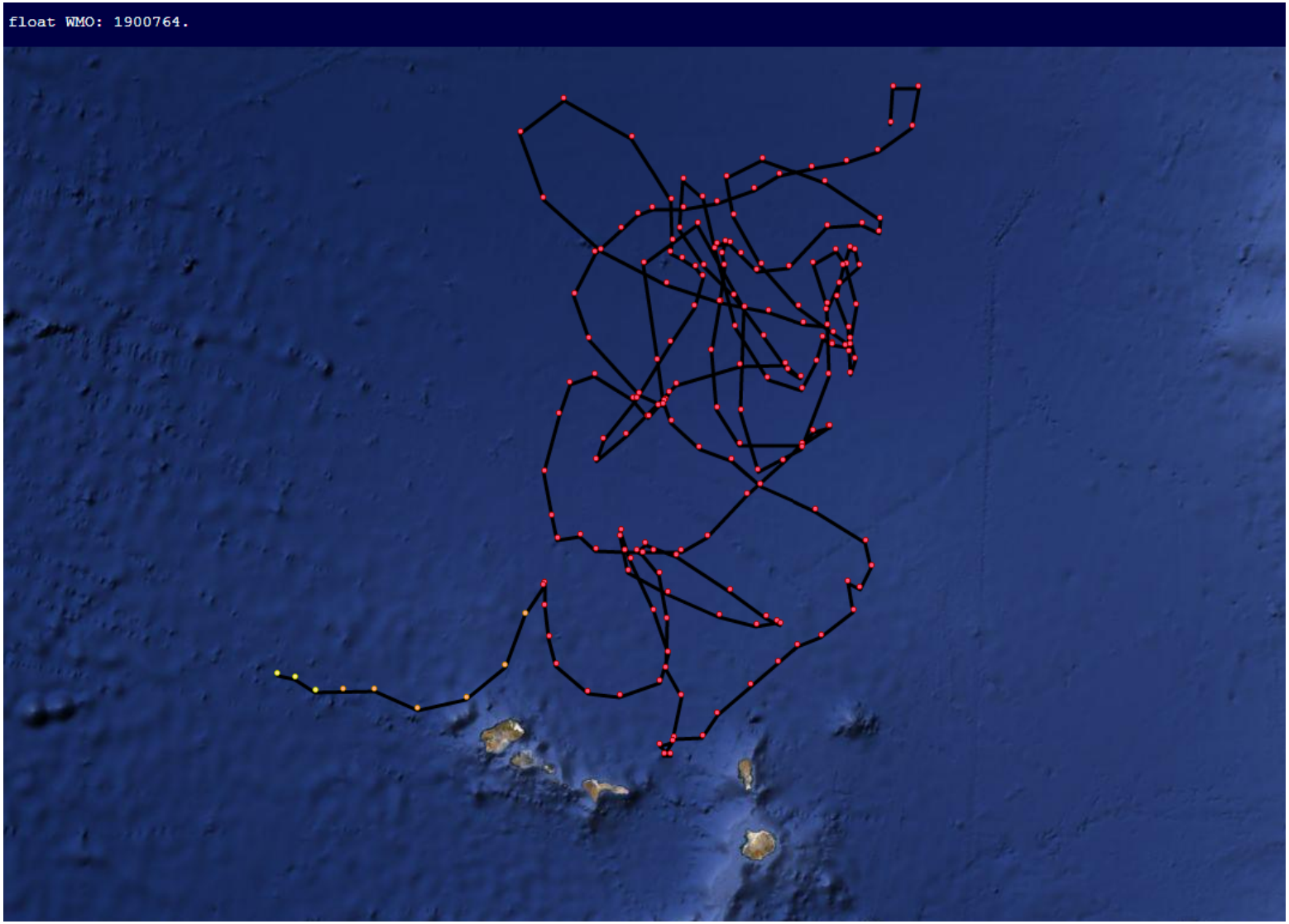


Location of collection: Equatorial Atlantic, near Africa

Canary Current dominates at surface, flows southwest



Location of collection: Near Cape Verde Islands



Models to be compared: Linear regression to predict Salinity

Two competitive models

Table: **salinity** $= f($**temperature**, **pressure**, **dM**, **latitude**, ***intercept***$)$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.63026	0.04194	801.87096	0.00000
temperature	0.11905	0.00082	145.13979	0.00000
pressure	0.00014	0.00001	11.84308	0.00000
.dM	-0.00936	0.00119	-7.85204	0.00000
latitude	0.02709	0.00189	14.31581	0.00000

Table: **salinity** $= f($**temperature**, **pressure**, **dM**, **latitude**$)$

	Estimate	Std. Error	t value	Pr(>|t|)
temperature	0.35584	0.00599	59.41448	0.00000
pressure	0.00342	0.00009	40.10497	0.00000
.dM	0.08848	0.00928	9.53769	0.00000
latitude	1.41079	0.00608	231.92096	0.00000

Models to be compared: Linear regression to predict Salinity

Two more competitive models

Table: **salinity** $= f($**temperature**, **pressure**, **dM**, ***intercept***$)$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.17769	0.01739	1965.64667	0.00000
temperature	0.11951	0.00083	144.44907	0.00000
pressure	0.00014	0.00001	12.19309	0.00000
.dM	-0.00969	0.00120	-8.05910	0.00000

Table: **salinity** $= f($**temperature**, **pressure**, **dM**$)$

	Estimate	Std. Error	t value	Pr(>|t|)
temperature	1.68979	0.00410	411.86732	0.00000
pressure	0.02155	0.00008	258.82884	0.00000
.dM	0.47638	0.02243	21.24306	0.00000

Models to be compared: Linear regression to predict Salinity

Still two more competitive models

Table: **salinity** $= f($**temperature**, **pressure**, ***intercept***$)$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.14888	0.01707	2000.88558	0.00000
temperature	0.11951	0.00083	144.02016	0.00000
pressure	0.00014	0.00001	12.17462	0.00000

Table: **salinity** $= f($**temperature**, **pressure**$)$

	Estimate	Std. Error	t value	Pr(>|t|)
temperature	1.75765	0.00263	668.68046	0.00000
pressure	0.02246	0.00007	308.89198	0.00000

Split-Half cross-validation example

Also called “Hold Out cross-validation” or “Split-Sample cross-validation”

Table: Split-half cross-validation summary

	MSE
temperature, pressure, dM, latitude; with intercept	0.217849
temperature, pressure, dM; with intercept	0.178725
temperature, pressure; with intercept	0.133647
temperature, pressure, dM, latitude; without intercept	10.852694
temperature, pressure, dM; without intercept	46.900654
temperature, pressure; without intercept	33.191302

K-fold Cross-validation example

Table: 10-fold cross-validation summary

	MSE
temperature, pressure, dM, latitude; with intercept	0.208880
temperature, pressure, dM; with intercept	0.210792
temperature, pressure; with intercept	0.211403
temperature, pressure, dM, latitude; without intercept	1.634176
temperature, pressure, dM; without intercept	4.016313
temperature, pressure; without intercept	4.099267

Bootstrap validation instead

Table: Bootstrap Validation summary (100 resamplings, Efron & Gong Bootstrap)

	MSE
temperature, pressure, dM, latitude; with intercept	0.043645
temperature, pressure, dM; with intercept	0.044209
temperature, pressure; with intercept	0.044584
temperature, pressure, dM, latitude; without intercept	0.043646
temperature, pressure, dM; without intercept	0.044245
temperature, pressure; without intercept	0.044723

Table: Bootstrap Validation summary (100 resamplings, 0.632+ Bootstrap)

	MSE
temperature, pressure, dM, latitude; with intercept	0.043482
temperature, pressure, dM; with intercept	0.044507
temperature, pressure; with intercept	0.044618
temperature, pressure, dM, latitude; without intercept	0.043574
temperature, pressure, dM; without intercept	0.044359
temperature, pressure; without intercept	0.044672

Model Selection and Comparison: Information Criteria

Something to be taught another day ... lecture on *predictive inference*

Table: Information Criteria Judging Quality of Several Models

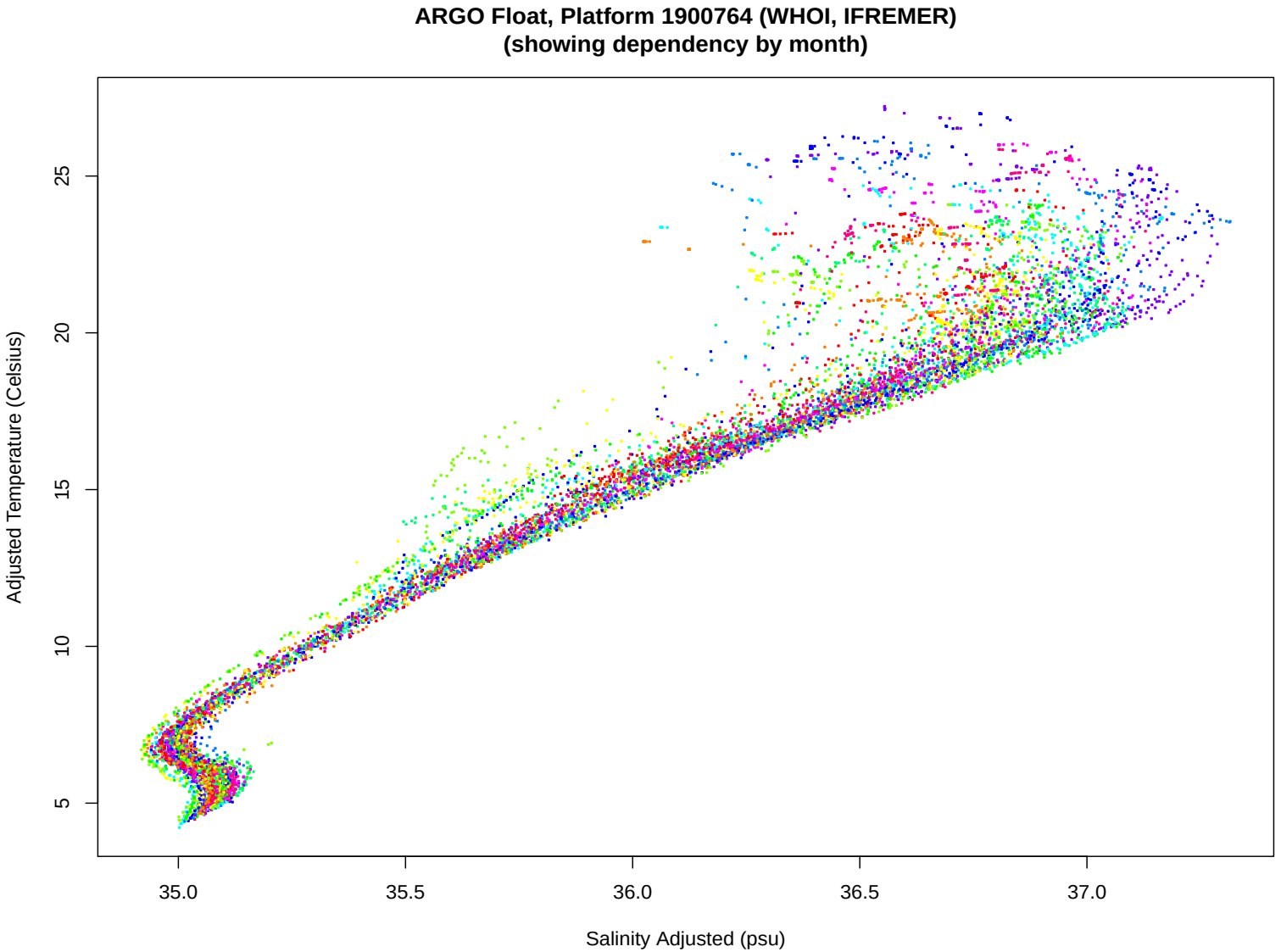
	AIC	AICc	BIC	R^2
temperature, pressure, dM, latitude; with intercept	-3146.61	-3146.60	-3102.96	0.91
temperature, pressure, dM; with intercept	-2945.52	-2945.51	-2909.14	0.91
temperature, pressure; with intercept	-2882.74	-2882.74	-2853.64	0.91
temperature, pressure, dM, latitude; without intercept	0.00	0.00	0.00	1.00
temperature, pressure, dM; without intercept	0.00	0.00	0.00	0.99
temperature, pressure; without intercept	0.00	0.00	0.00	0.99

Table: Relative Importance of Various Attributes in Several Models

	temperature	pressure	.dM	latitude
temperature+pressure+dM+latitude+*intercept* model	0.5917	0.4046	0.0005	0.0032
temperature+pressure+dM+*intercept* model	0.5939	0.4056	0.0005	0.0000
temperature+pressure+*intercept* model	0.5942	0.4058	0.0000	0.0000

Variation by season: Salinity as predicted by temperature, pressure, latitude, month

ARGO float, platform 1900764 (WHOI, IFREMER)



Computation

- **R**
- *SEQBoot* special purpose and free
- Nothing in GNU Scientific Library, unfortunately.
- For *non-delicate inferences*, can simply write the sampling yourself.
- Kleiner, Talwalkar, Sarkar, Jordan “Big data bootstrap”, 2012

When do these fail?

- Fail principally when data are co-correlated or *dependent*
- Special techniques for *dependent data*: Next lecture!
- Fail when trying to characterize extremes of distributions like minimum, maximum: Insufficient number of samples
- Deteriorates when *sampling distribution* of population is *not smooth*, in a formal sense

Bibliography

- M. R. Chernick, *Bootstrap Methods: A Guide for Practitioners and Researches*, 2^{nd} edition, 2008, John Wiley & Sons.
- A. C. Davison, D. V. Hinkley, *Bootstrap Methods and their Application*, first published 1997, 11^{th} printing, 2009, Cambridge University Press.
- R. Durrett, *Probability: Theory and Examples*, 4^{th} edition, Cambridge University Press, 2010.
- B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap*, 1993, Chapman & Hall/CRC.
- P. I. Good, *Resampling Methods – A Practical Guide to Data Analysis*, 2006.
- B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, 1982.
- P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, 1992.
- K. Singh, G. J. Babu, "On the asymptotic optimality of the bootstrap", *Scandanavian Journal of Statistics*, 17, 1990, 1-9.
- A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, "The Big Data Bootstrap", in J. Langford and J. Pineau (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, UK, 2012.
- A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, "A scalable bootstrap for massive data", `http://arxiv.org/abs/1112.5016`, last accessed 16^{th} November 2012.
- B. Efron, R. Tibshirani, "Improvements on cross-validation: The 0.632+ Bootstrap method", *JASA*, 92(438), June 1997, 548-560.