Biostatistics-Lecture 6

Estimation, confidence interval and hypothesis testing

Ruibin Xi

Peking University

School of Mathematical Sciences

Some Results in Probability (1)

- Suppose that X, Y are independent ($X \perp Y$)
  - E(cX) = ? (c is a constant)
  - E(X+Y) = ?
  - Var(cX) = ?
  - Var(X+Y) = ?
- Suppose $X_1 \cdots X_n$ are mutually independent identically distributed (i.i.d.)
  - $E(\overline{X}_n) = ?$
  - $Var(\overline{X}_n) = ?$

Some Results in Probability (2)

- The Law of Large Number (LLN)

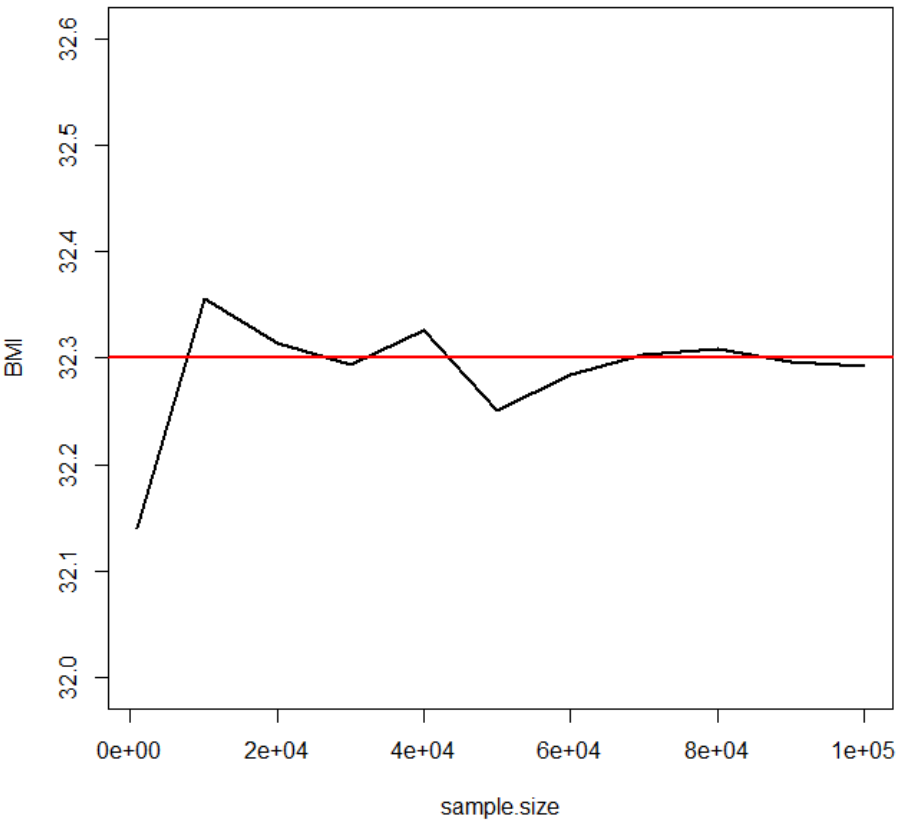
The **Law of Large Numbers (LLN)** indicates that (under some general conditions such as independence of observations) the sample mean converges to the population mean ($\bar{X}_n \to \mu$) as the sample size n increases ($n \to \infty$). Informally, this means that the difference between the sample mean and the population mean tends to become smaller and smaller as we increase the sample size. The LLN provides a theoretical justification for the use of sample mean as an estimator for the population mean.

- Assume BMI follows a normal distribution with mean 32.3 and sd 6.13

Some Results in Probability (2)

- The Law of Large Number (LLN)

The **Law of Larg**eneral con-
ditions such as inc onverges to
the population me $\rightarrow \infty$). In-
formally, this mea nd the pop-
ulation mean tend the sample
size. The LLN pro mple mean
as an estimator fo



- Assume Bl ution
with mear

Some Results in Probability (3)

- The Central Limit Theorem (CLT)

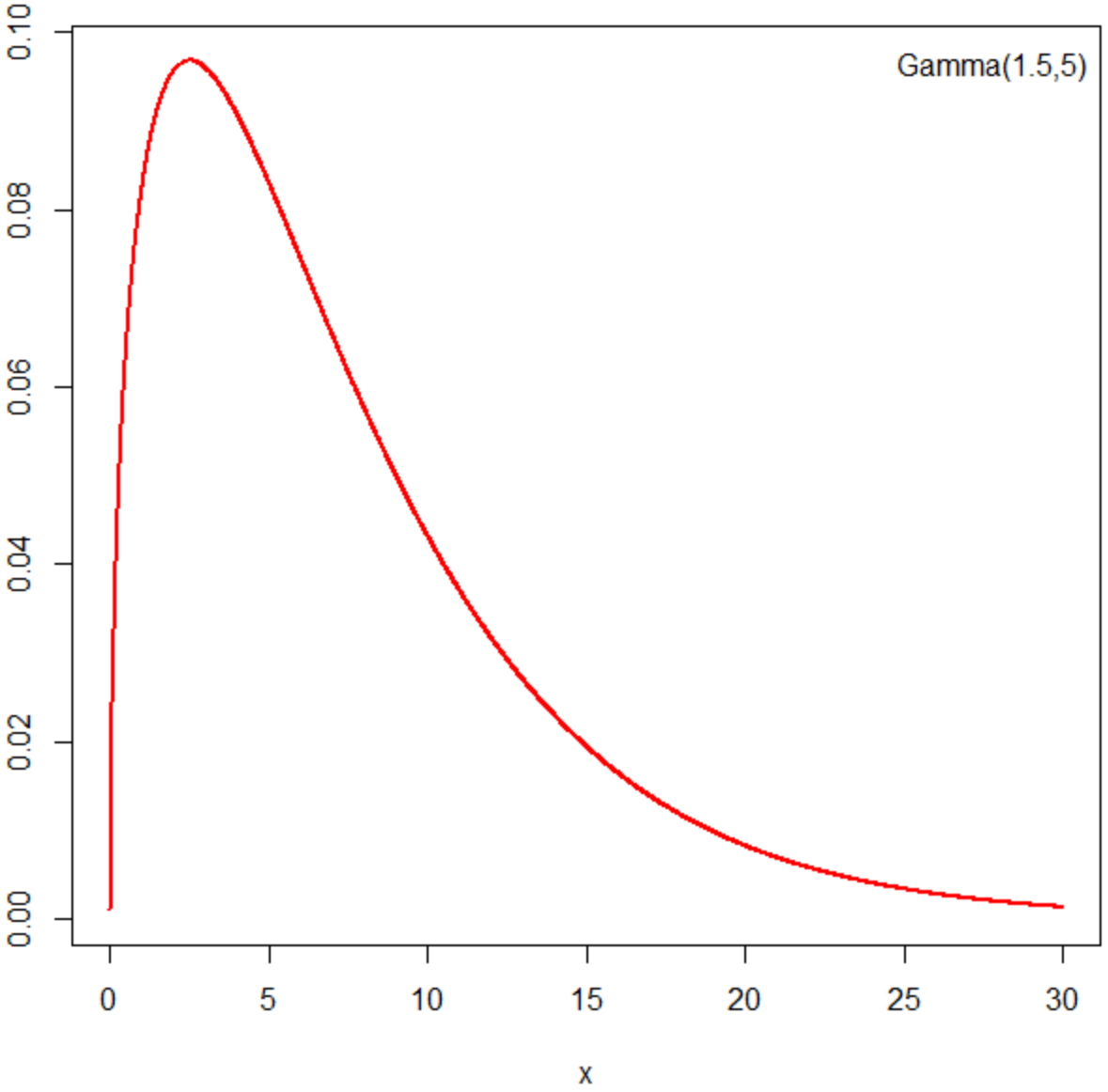
For large sample sizes, the CLT indicates that (under certain conditions such as independence of observations) if the random variable X has the population mean μ and the population variance σ^2, then the sampling distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n:

$$\bar{X} \sim N\left(\mu,\ \sigma^2/n\right).$$

S

- The

For large ns such
as indepe pulation
mean μ on of $\bar{X}$
is approx

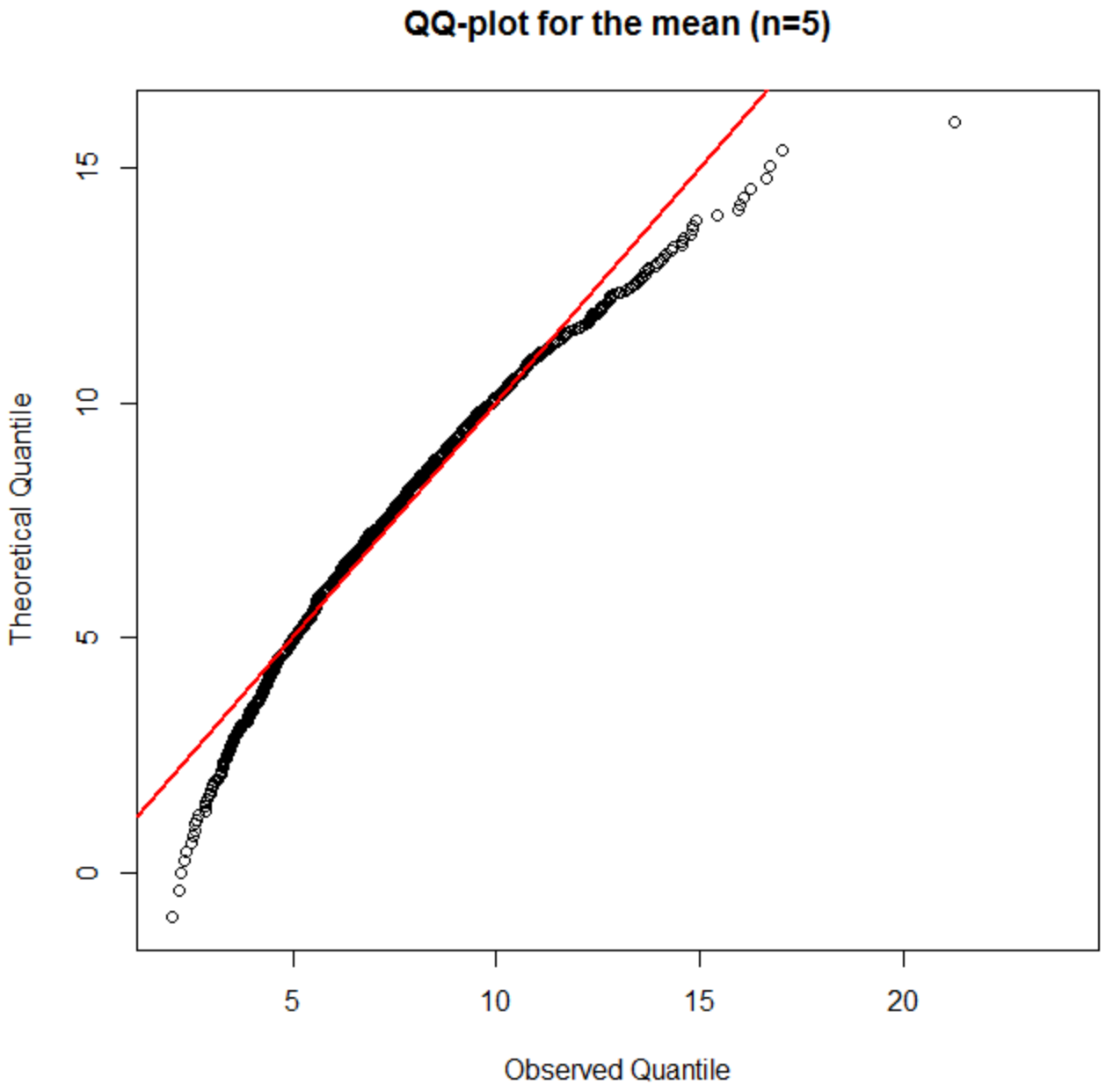


S

3)

- The

For large ns such
as indepe pulation
mean μ on of $\bar{X}$
is approx



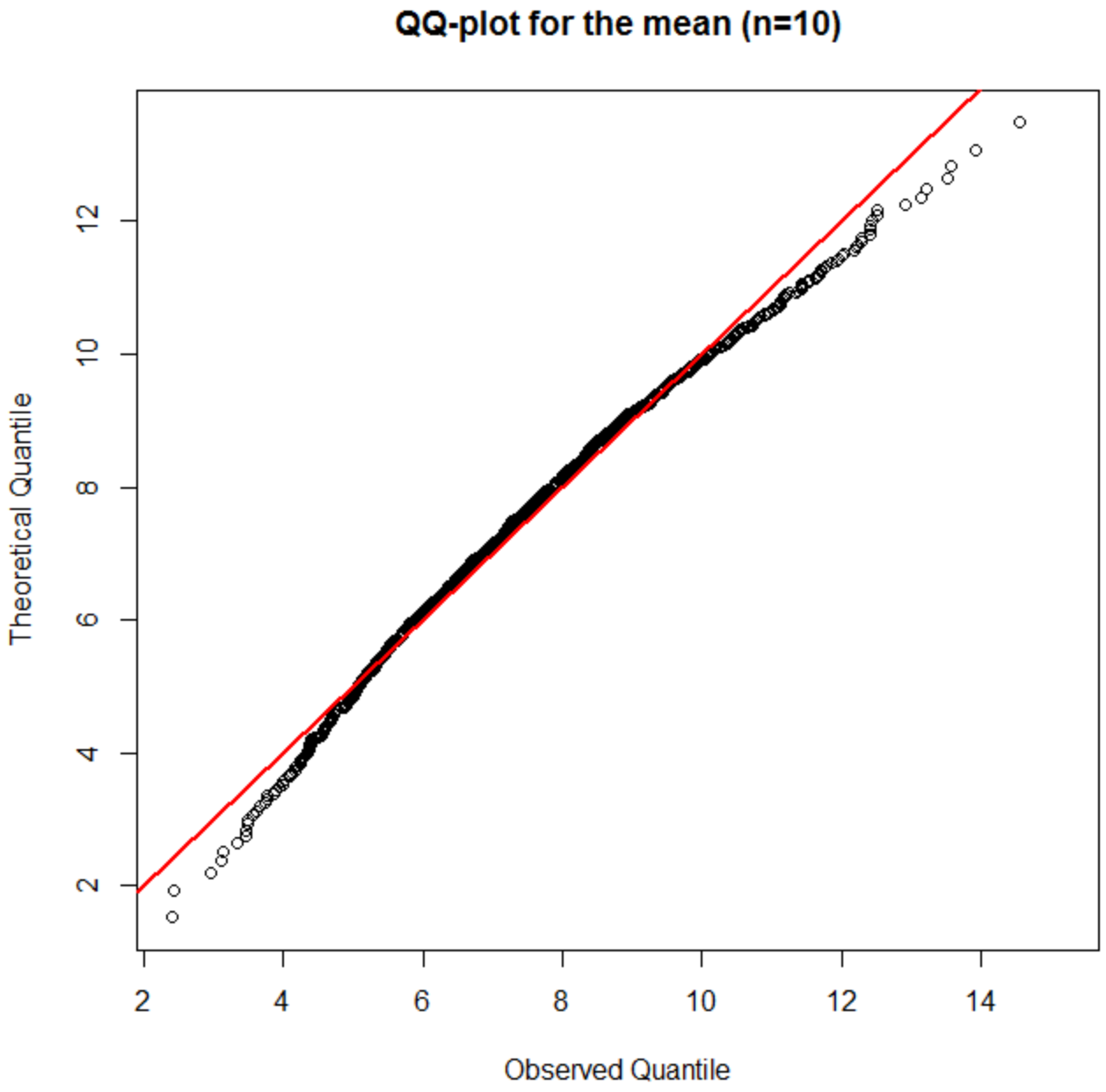
S

)

- The

For large
as indep
mean μ
is approx

ns such
ɔulation
ɔn of $\bar{X}$

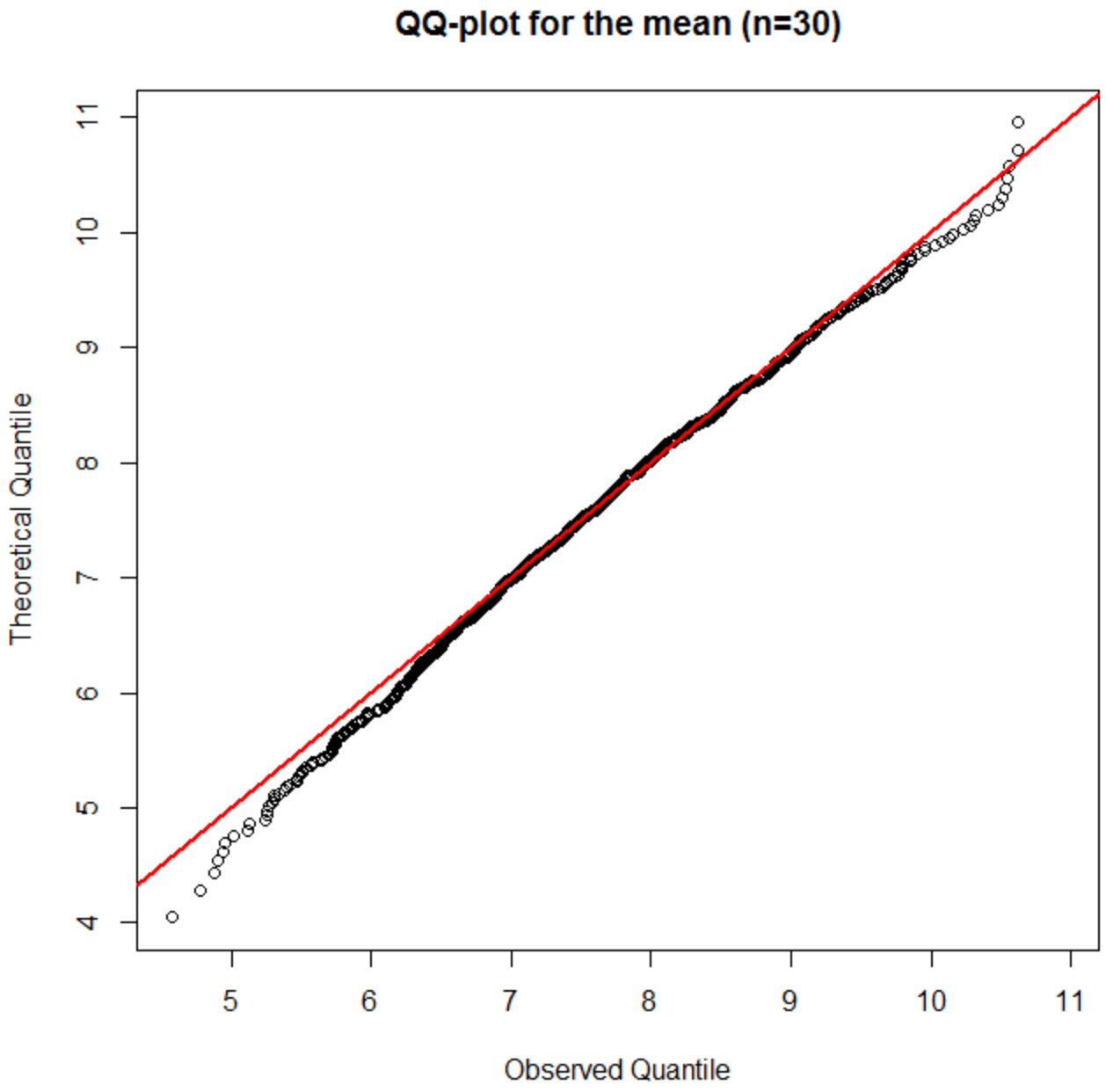


S

- The

For large
as indep
mean μ
is approx

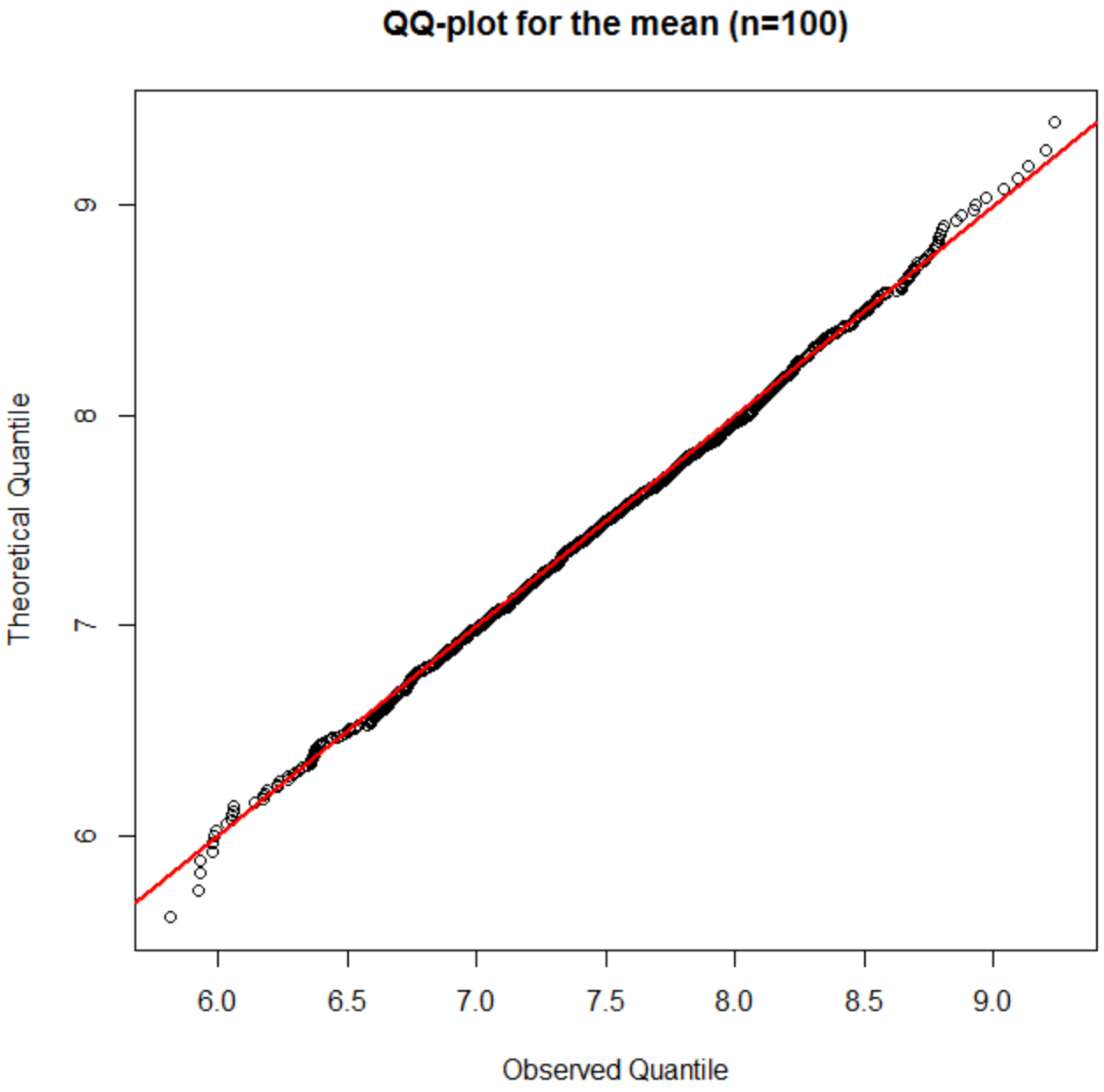
ns such
ɔulation
ɔn of $\bar{X}$



S

- The

For large ns such
as indepe pulation
mean μ on of $\bar{X}$
is approx



Statistical Inference

- Draw conclusions about a population from a sample
- Two approaches
  - Estimation
  - Hypothesis testing

Estimation

- Point estimation—summary statistics from sample to give an estimate of the true population parameter

$$\bar{X} \to \mu$$

$$s \to \sigma$$

  - The LLN implies that when n is large, these should be close to the true parameter values
  - These estimates are random
- Confidence intervals (CI): indicate the variability of point estimates from sample to sample

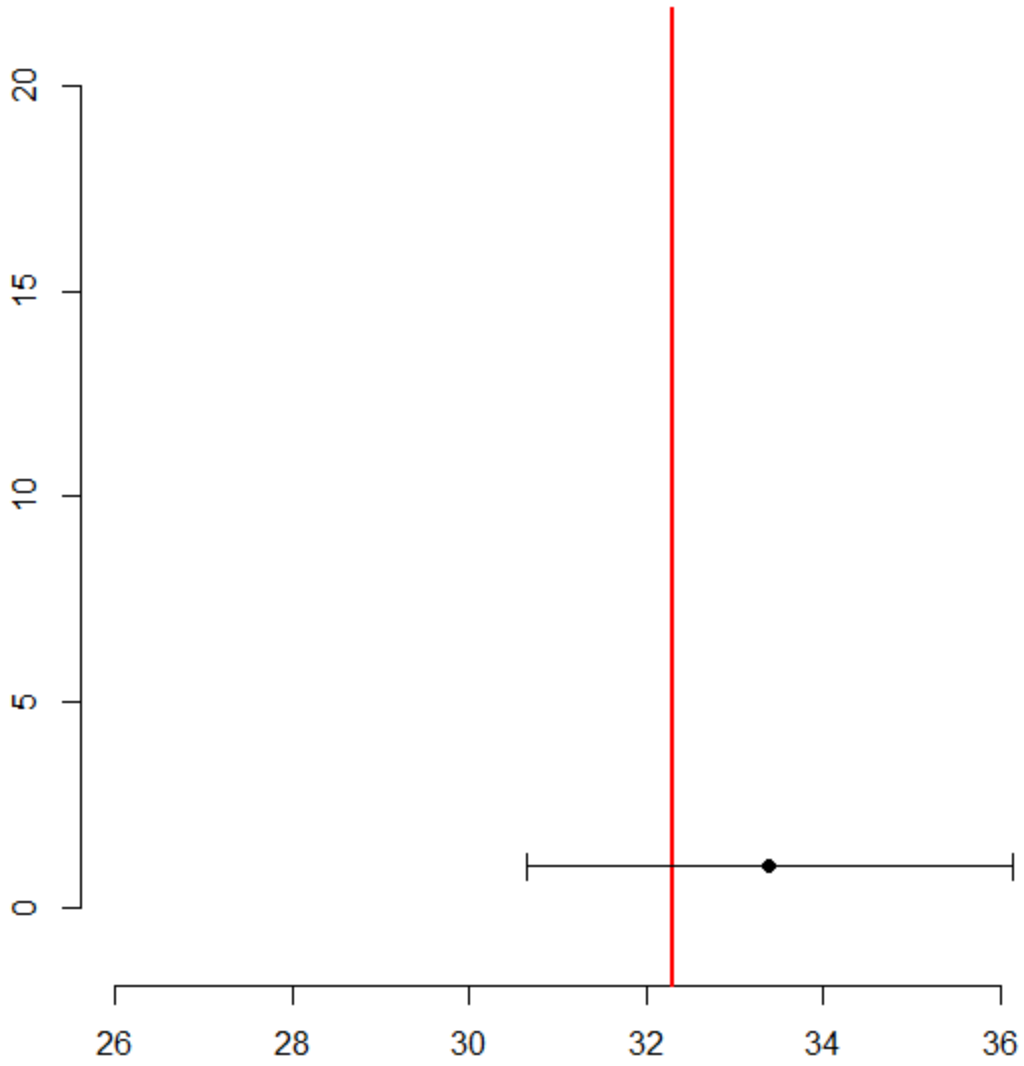
Confidence interval

- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ (σ is known)
  - $P(|\bar{X}_n - \mu| \leq \frac{2\sigma}{\sqrt{n}}) \approx 0.95$
  - Confidence interval of level 95% $\left[\bar{X}_n - \frac{2\sigma}{\sqrt{n}}, \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right]$
  - Repeatedly construct the confidence interval, 95% of the time, they will cover μ
  - In the BMI example, μ=32.3, σ=6.13, n = 20

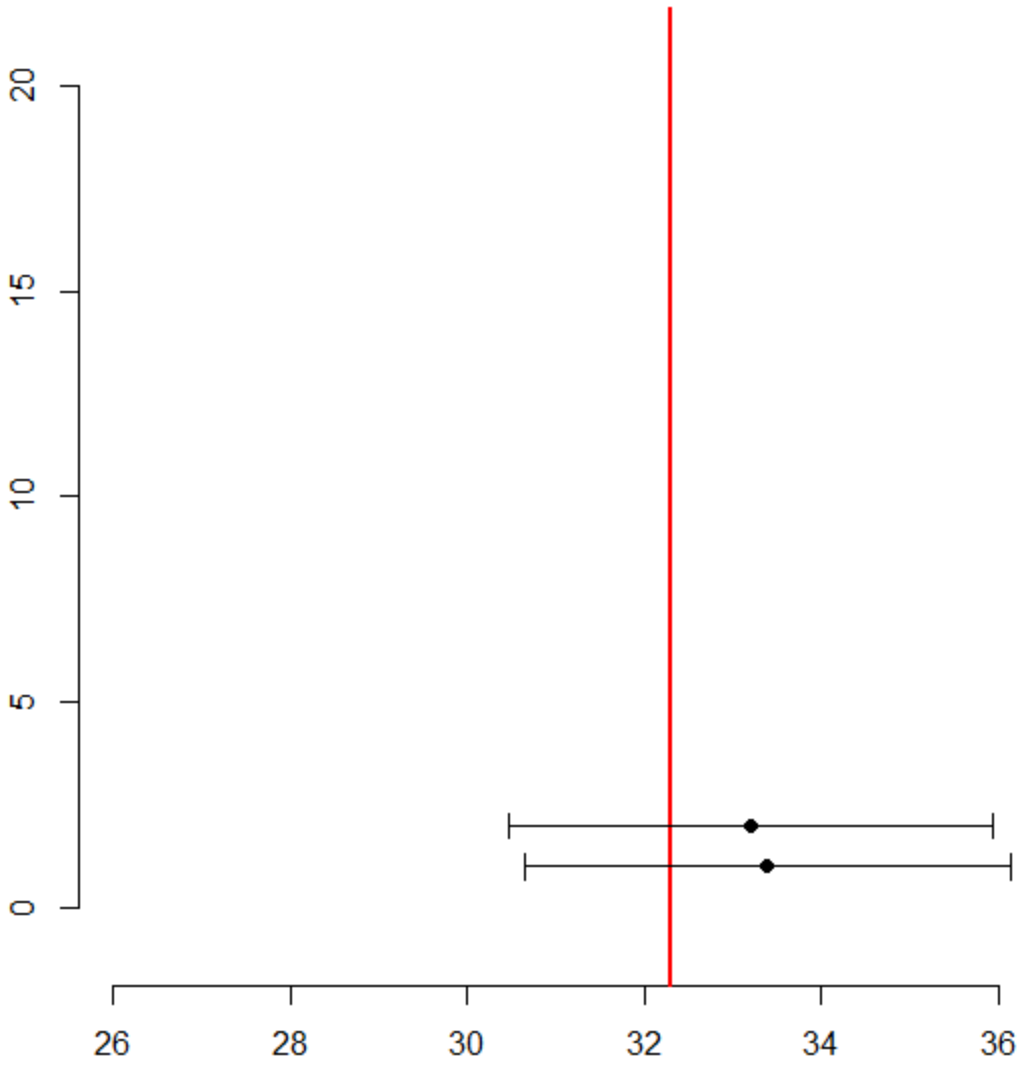
Confidence interval

- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ (σ is known)
  - $P(|\bar{X}_n - \mu| \leq \frac{2\sigma}{\sqrt{n}}) \approx 0.95$
  - Confidence interval $\left[\bar{X}_n - \frac{2\sigma}{\sqrt{n}}, \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right]$
  - Repeatedly construct the confidence interval, 95% of the time, they will cover μ
  - In the BMI example, μ=32.3, σ=6.13, n = 20

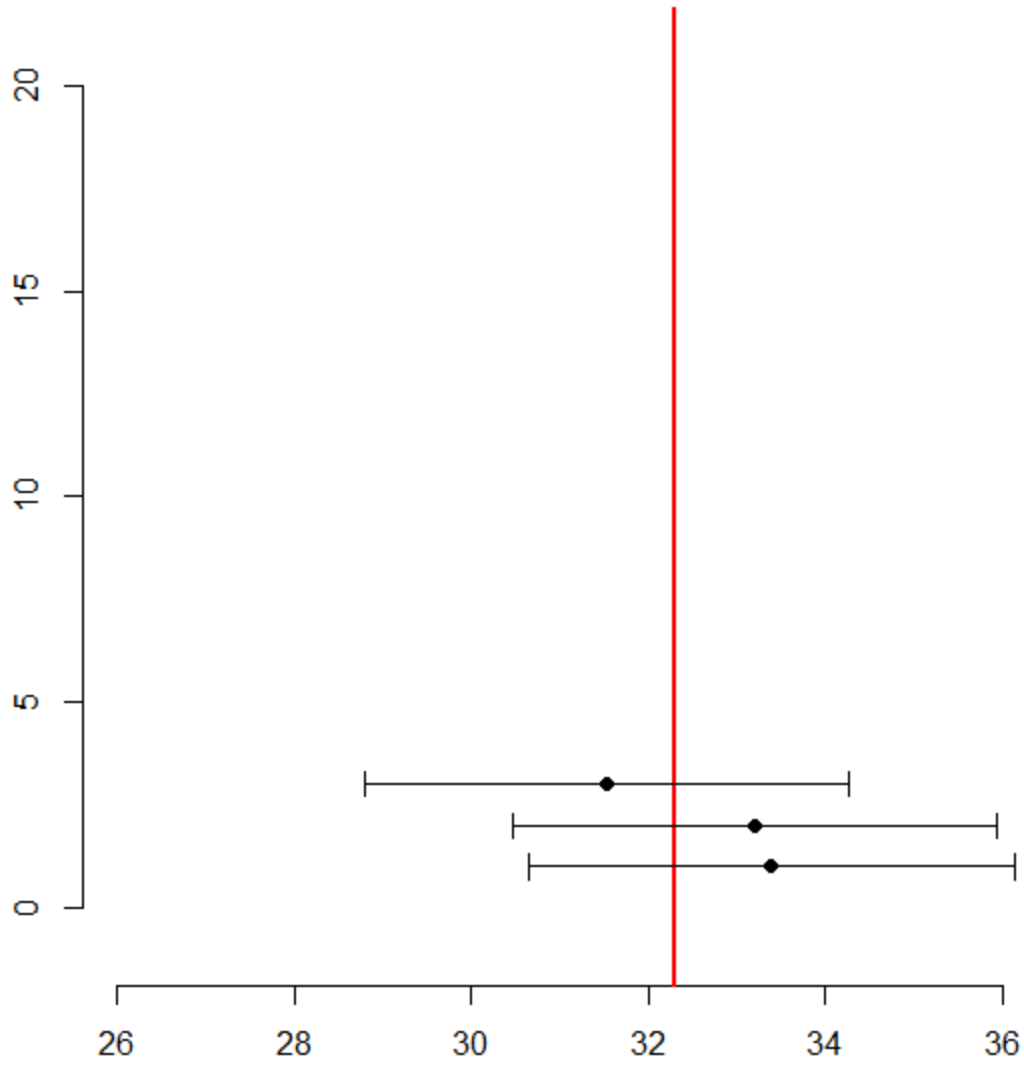
- Ass
kno
  -
  - (
  -
  - g



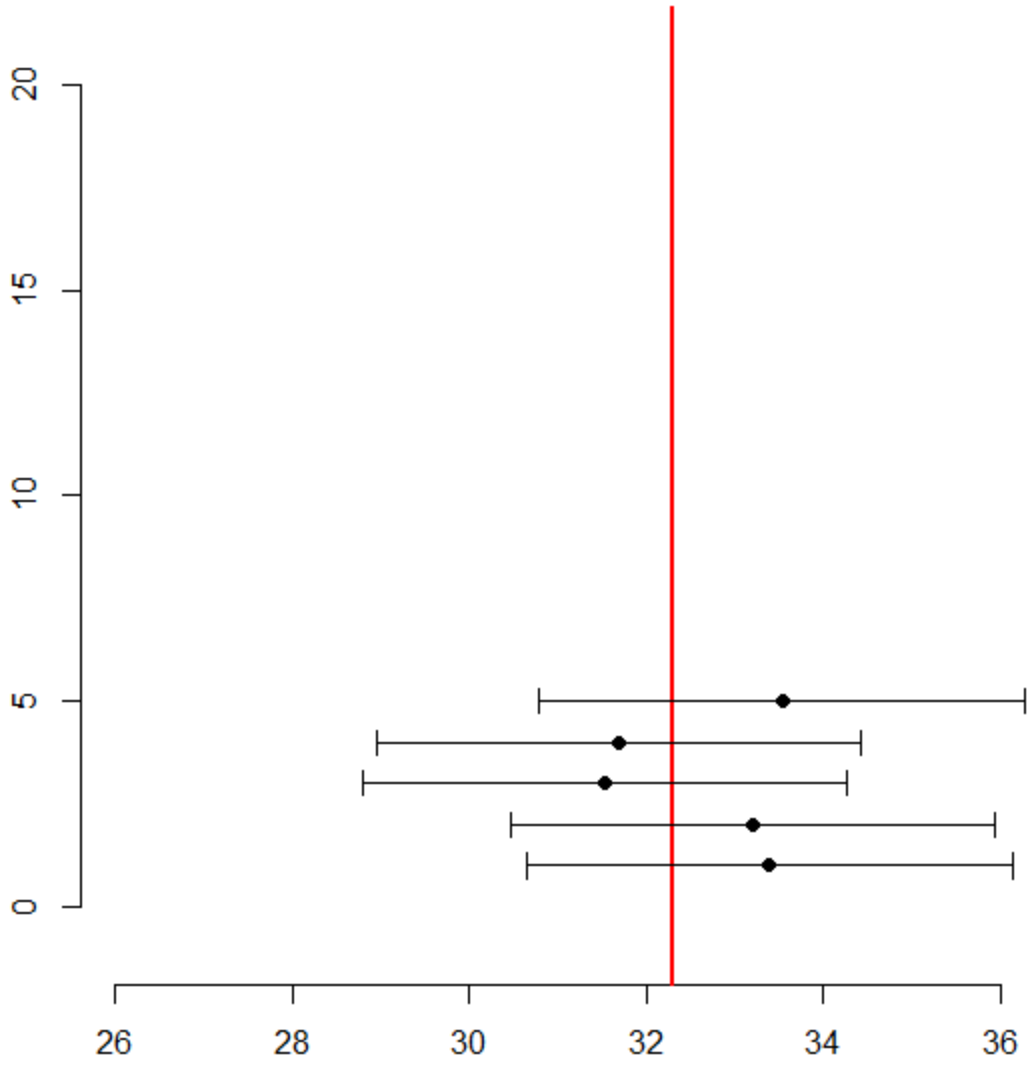
- Ass
  kno
  -
  -
  -
  - g



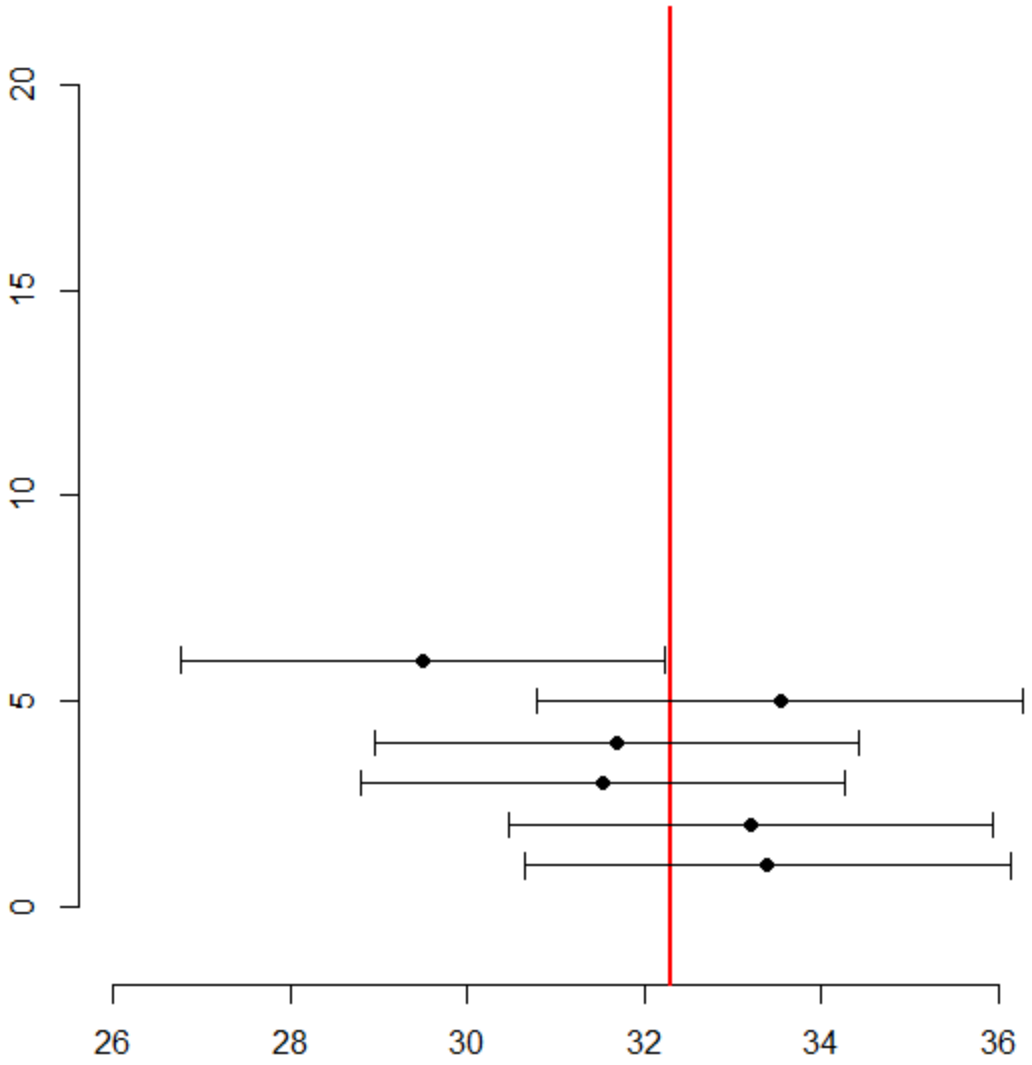
- Ass
kno
  -
  - (
  -
  - g



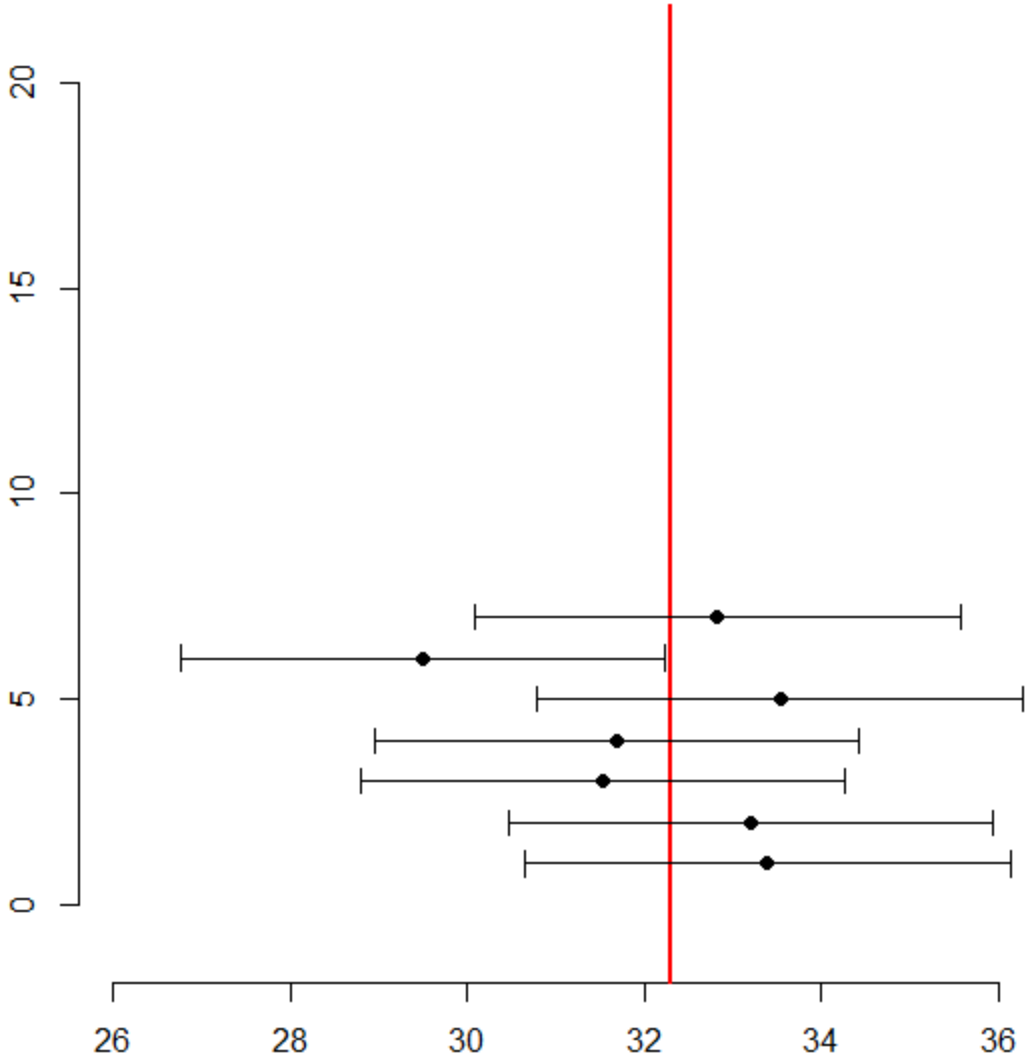
- Ass
  kno
  -
  -
  -
  - g



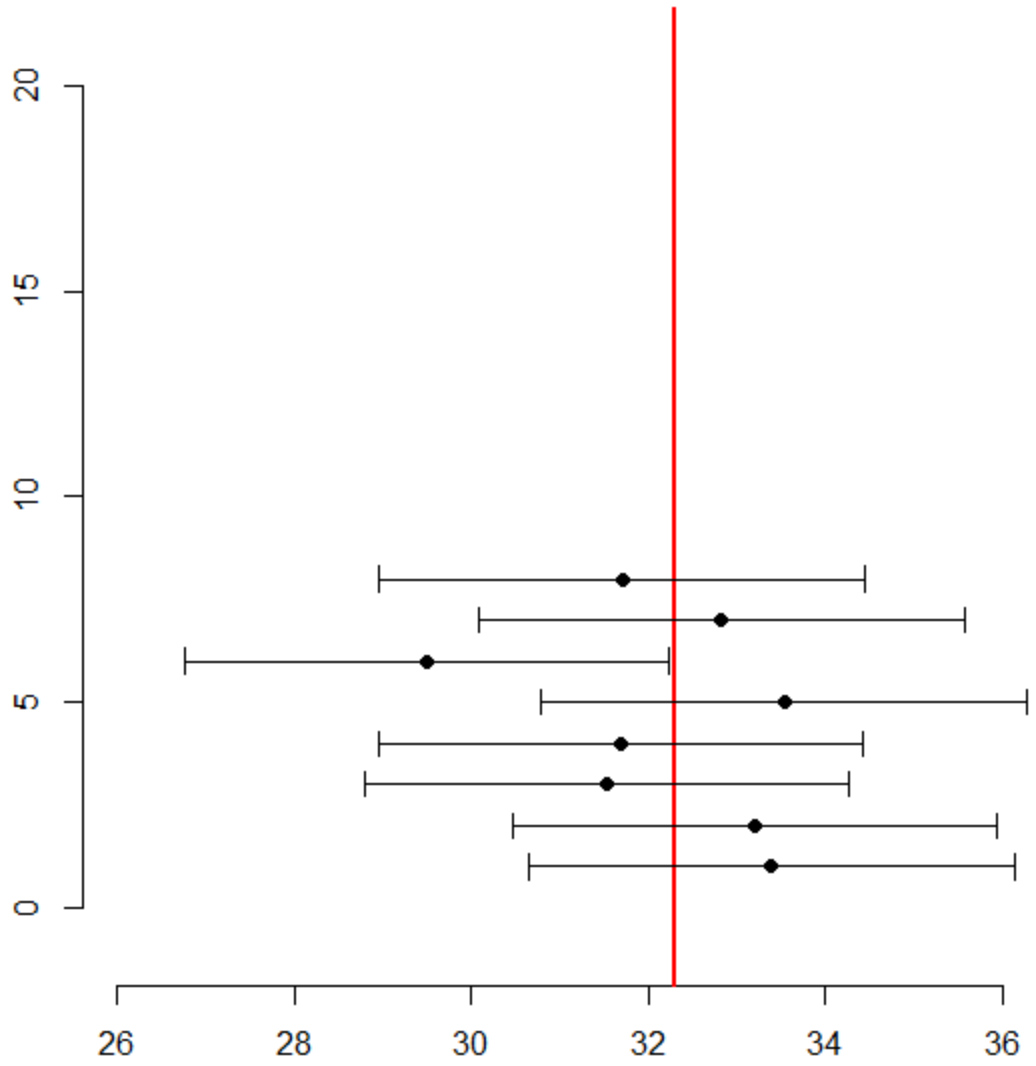
- Ass
  kno
  -
  -
  -
  - g



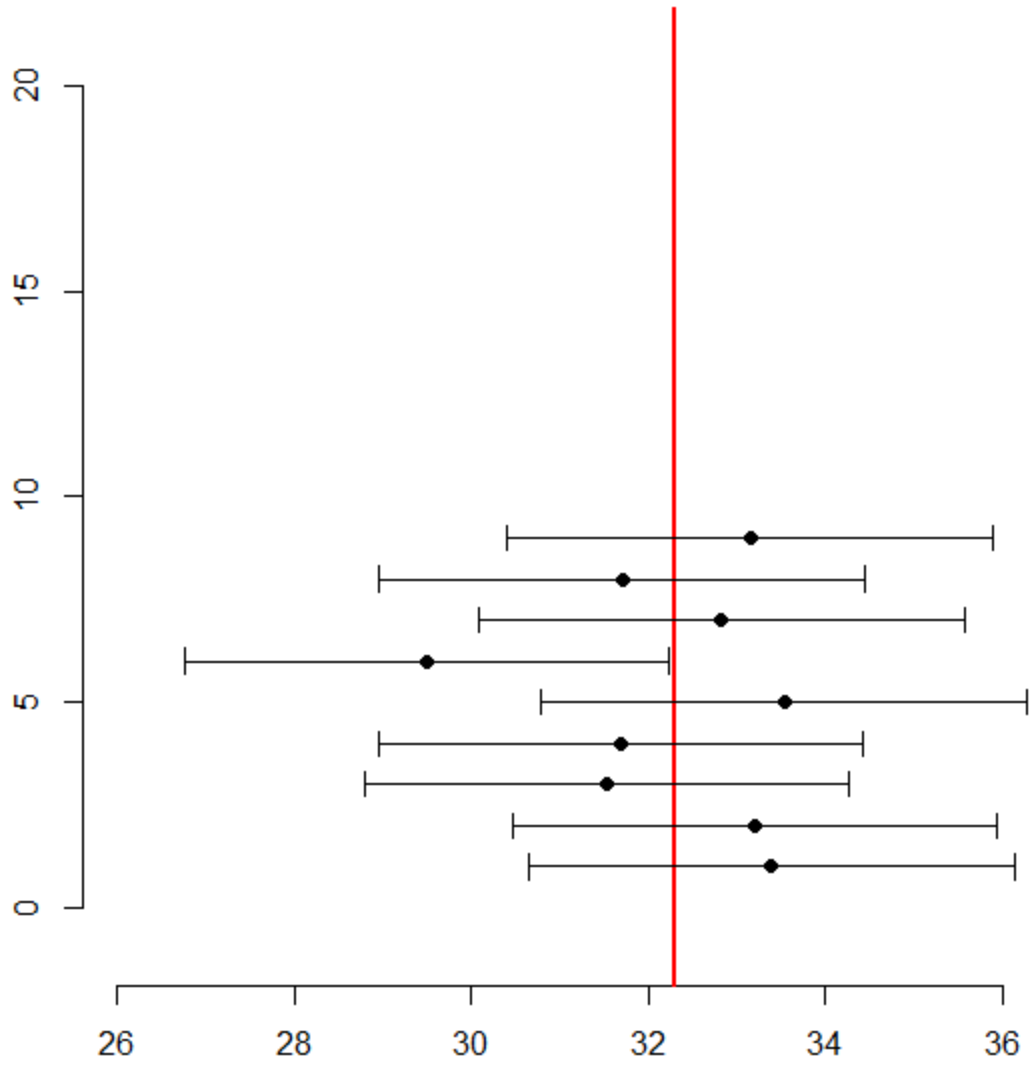
- Ass
kno
  -
  -
  - g



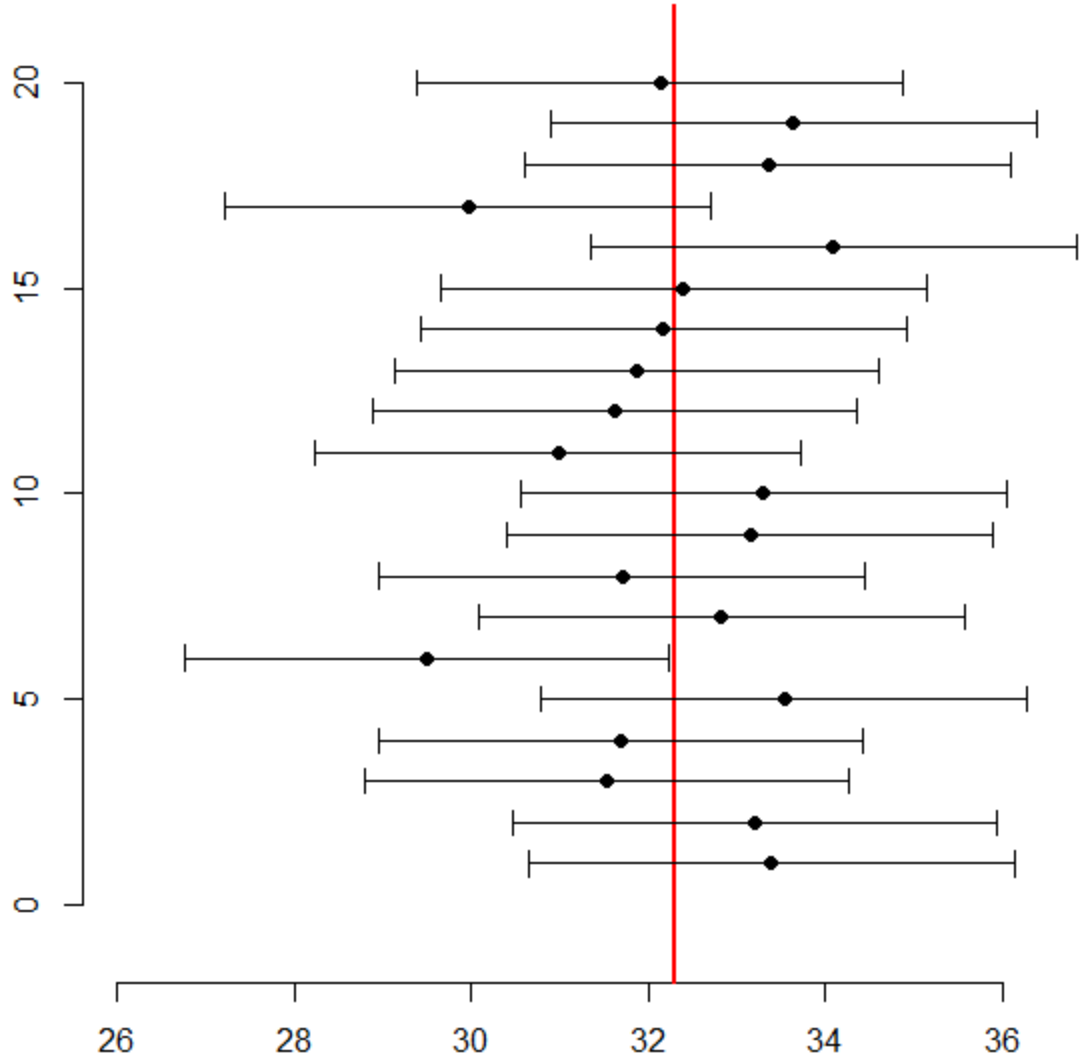
- Ass
kno
  -
  -
  - g



- Ass
kno
  -
  -
  - g



- Ass
  kno
  -
  -
  -
  - g



Confidence Interval for the Mean

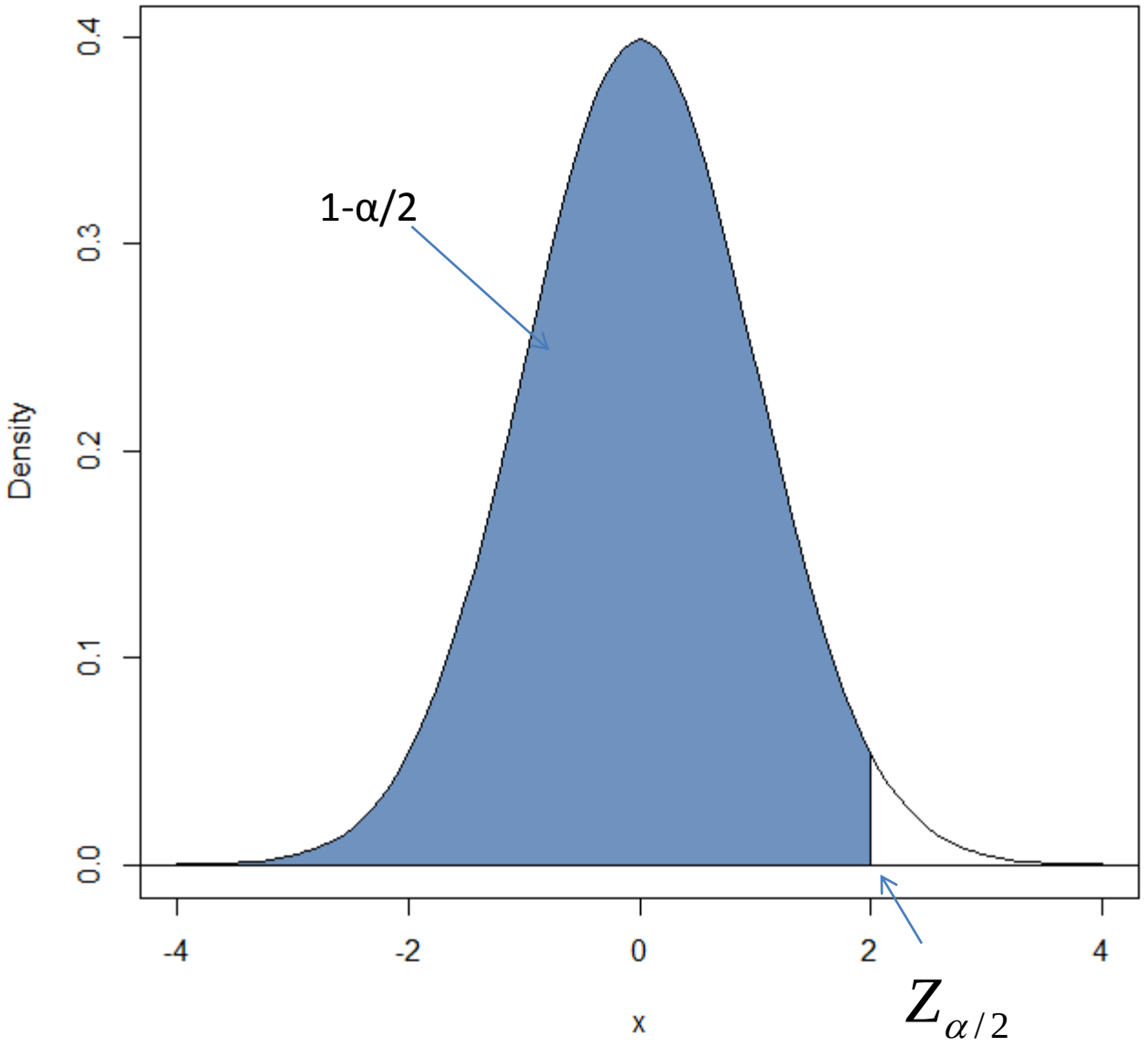
- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $(\bar{X}_n - \mu) \Big/ \frac{\sigma}{\sqrt{n}} \sim N(0,1)$ (σ is known)
  - $P(|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 1 - \alpha$
  - Confidence interval of level 1-α $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right]$

Confidence Interval for the Mean

- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ (σ is known)
  - $P(|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 1 - \alpha$ (1-α/2)
  - Confidence interval of level 1-α $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right]$

Co an

- Ass
  kno
  - *P*
  - (



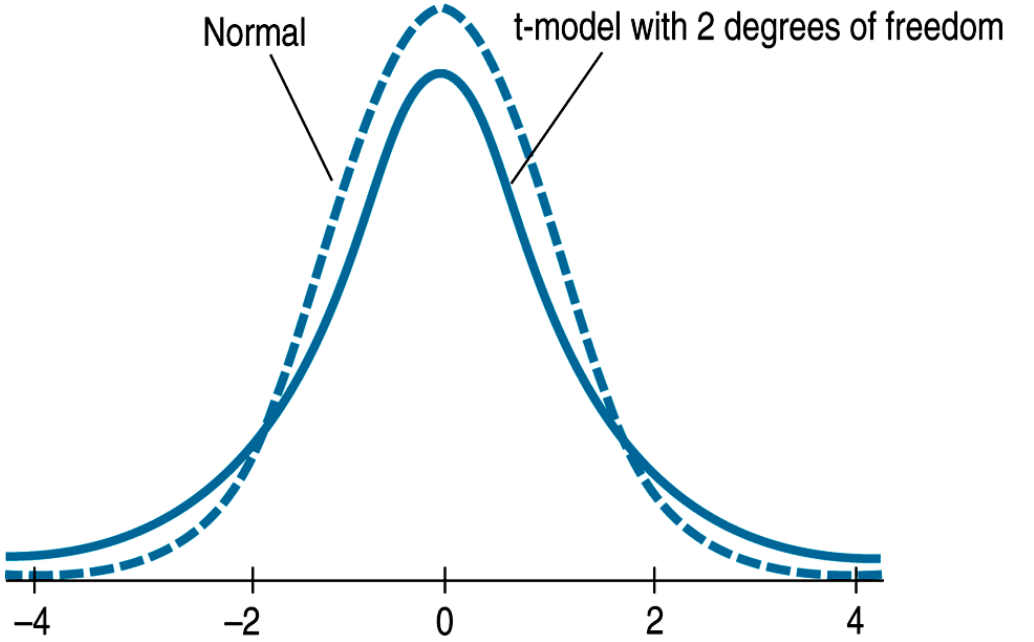
$\cdot \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$]

Confidence Interval for the Mean

- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $(\bar{X}_n - \mu) \Big/ \frac{\sigma}{\sqrt{n}} \sim N(0,1)$ (σ is known)
  - $P(|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 1 - \alpha$
  - Confidence interval of level 1-α $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right]$
- What if σ is unknown?
  - t-statistics!

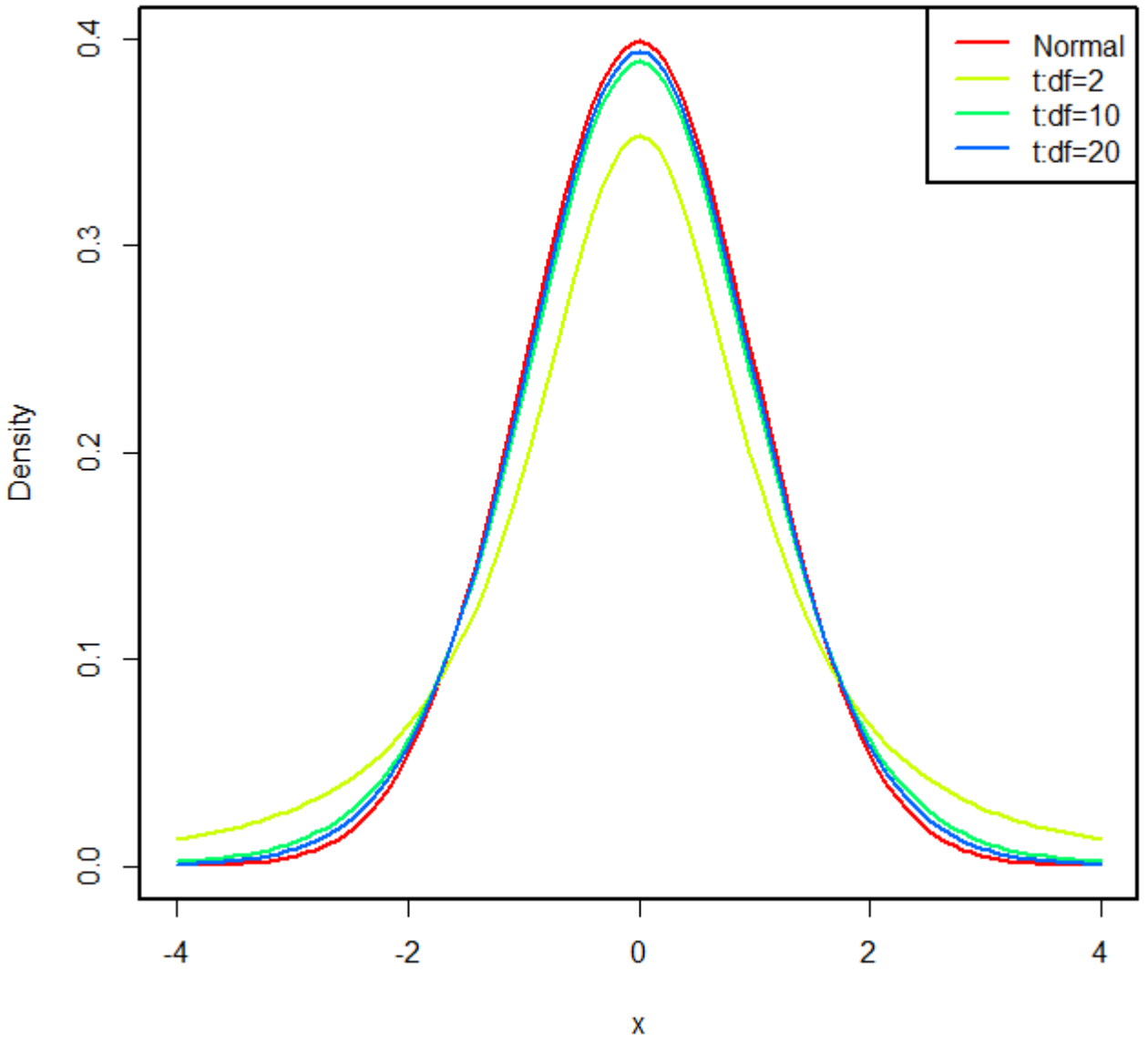
Confidence Interval for the Mean

- Assume $X_1 \cdots X_n \sim N(\mu, \sigma^2)$, then $(\bar{X}_n - \mu) \Big/ \frac{\sigma}{\sqrt{n}} \sim N(0,1)$
  - $S_n^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \to \sigma^2$ by the LLN.
  - Replace σ^2 by $S_n^{\,2}$, then $(\bar{X}_n - \mu) \Big/ \frac{S_n}{\sqrt{n}} \sim t_{n-1}$ ← Standard error (SE)
  - Confidence interval of level 1-α $\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{\alpha/2}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{\alpha/2}\right]$



Co ean

- As



$$= t_{\alpha/2}$$

Confidence Interval for the Mean

- Measure serum cholesterol (血清胆固醇) in 100 adults

$$\bar{x} = 6.7 mmol/L$$

$$s = 1.2 mmol/L$$

- Construct a 95% CI for the mean serum cholesterol based on t-distribution

$$6.7 \pm t_{99,0.975} \frac{1.2}{\sqrt{100}} = 6.7 \pm 1.98 \times \frac{1.2}{\sqrt{100}} \quad [6.46,6.94]$$

- CI based on normal distribution

$$6.7 \pm 1.96 \times \frac{1.2}{\sqrt{100}} \quad [6.46,6.93]$$

Confidence interval based on the CLT

- Assume $X_1 \cdots X_n$ are i.i.d. random variable with population mean μ and population variance σ^2
  - Construct CI for μ?
  - From the CLT, approximately, $(\bar{X}_n - \mu) \Big/ \frac{\sigma}{\sqrt{n}} \sim N(0,1)$
  - From the LLN, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \to \sigma^2$
  - The asymptotic CI of level 1-α is $\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} Z_{\alpha/2}, \bar{X}_n + \frac{S_n}{\sqrt{n}} Z_{\alpha/2}\right]$

Confidence Interval for the proportions

- Telomerase
  - a ribonucleoprotein polymerase
  - maintains telomere ends by addition of the telomere repeat TTAGGG
  - usually suppressed in postnatal somatic cells
  - Cancer cells (~90%) often have increased telomerase activity, making them immortal (e.g. HeLa cells)
  - A subunit of telomerase is encode by the gene *TERT* (telomerase reverse transcriptase)

Confidence Interval for the proportions

- Huang et. al (2013) found that *TERT* promoter mutation is highly recurrent in human melanoma
  - 50 of 70 has the mutation
- Construct a 95% CI for the proportion (p) of melanoma genomes that has the *TERT* promoter mutation
  - From the data above, our estimate is $\hat{p} = \bar{x} = 50/70 = 0.714$
  - The standard error is $SE = \sqrt{\hat{p}(1-\hat{p})/n} = 0.054$
  - The CI is $[\hat{p} - 1.96 * SE, \hat{p} + 1.96 * SE] = [0.61, 0.82]$
- Note: to guarantee this approximation good, need p and 1-p ≥ 5/n

Hypothesis testing

- Scientific research often starts with a hypothesis
  - Aspirin can prevent heart attack
  - Imatinib can treat CML patient
  - TERT mutation can promote tumor progression
- Collect data and perform statistical analysis to see if the data support the hypothesis or not

Steps in hypothesis testing

- Step 1. state the hypothesis
  - Null hypothesis

    H_0: no different, effect is zero or no improvement
  - Alternative hypothesis

    H1: some different, effect is nonzero

    Directionality—one-tailed or two-tailed

    $\mu<$constant

    $\mu\neq$constant

Steps in hypothesis testing

- Step 2. choose appropriate statistics
  - Test statistics depends on your hypothesis
    - Comparing two means

      z-test or t-test
    - Test independence of two categorical variables

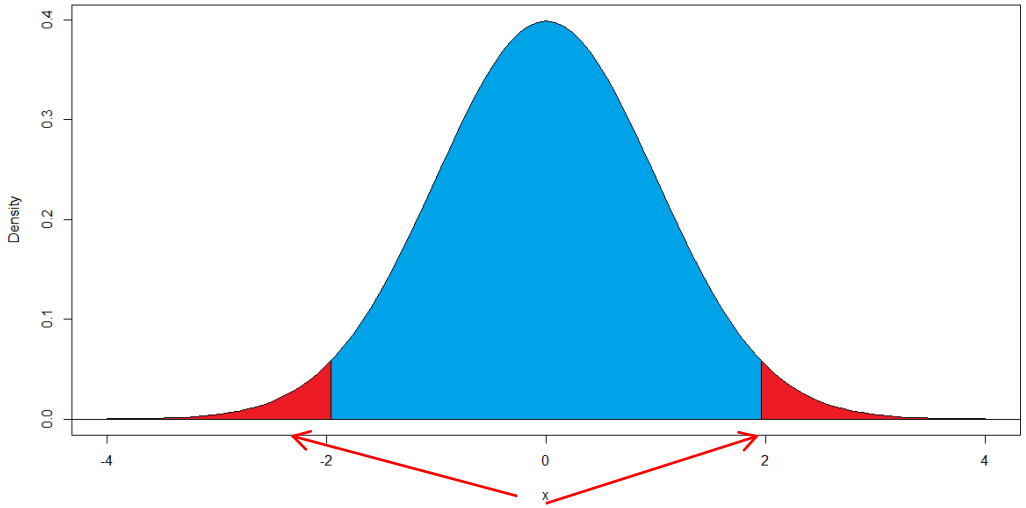
      Fisher's test or chi-square test

Steps in hypothesis testing

- Step 3. Choose the level of significance—α
  - How much confidence do you want in decision to reject the null hypothesis
  - α is also the type I error or false positive level
  - Typically 0.05 or 0.01

Steps in hypothesis testing

- Step 4. Determine the critical value of the test statistics that must be obtained to reject the null hypothesis under the significance level
  - Example—two-tailed 0.05 significance level for z-test



Rejection region

Steps in hypothesis testing

- Step 5. Calculate the test statistic
  - Example: t-statistic

$$t = \left(\overline{X}_n - \mu\right) \Big/ \frac{S_n}{\sqrt{n}}$$

- Step 6. Compare the test statistic to the critical value
  - If the test statistic is more extreme than the critical value, reject H_0

    DO NOT ACCEPT H_1
  - Otherwise, Do Not reject or Fail to reject H_0

    DO NOT ACCEPT H_0

Steps in hypothesis testing: an example

- Data Pima.tr in the MASS package
  - Data from Pima Indian heritage women living in USA (≥21) testing for diabetes
  - Question: Is the mean BMI of Pima Indian heritage women living in USA testing for diabetes is the same as the mean women BMI (26.5)
- Step 1. state the hypothesis
  - Let μ be the mean BMI of Pima Indian heritage women living in USA
  - H0: $\mu=26.5$; H1: $\mu\neq26.5$

Steps in hypothesis testing: an example

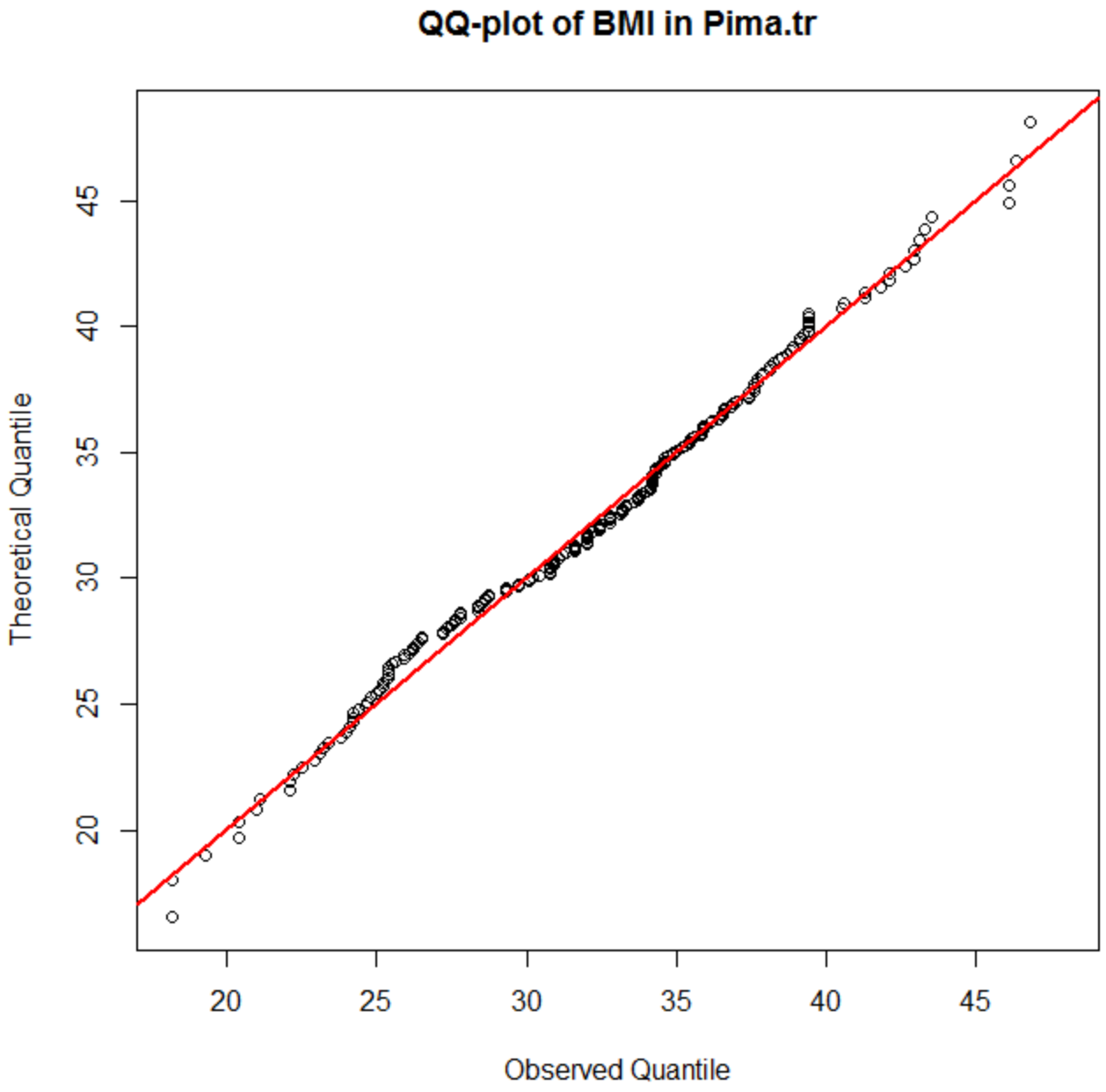
- Step 2. Choose appropriate test
  - Two-sided t-test
    - Hypotheses problem

      $\mu = \mu_0$; H1: $\mu \neq \mu_0$
    - Assumptions

      $X_1 \cdots X_n \sim N(\mu, \sigma^2)$ are independent, σ is unknown
    - Test statistic $T = \left(\overline{X}_n - \mu_0\right) \Big/ \frac{S_n}{\sqrt{n}}$ (under H0, follows t_{n-1})
    - Critical value

      $$P(|T| > C_{cri,\alpha}) = \alpha$$
  - Check if the test is appropriate

- Ste
  - T
  - C


QQ-plot of BMI in Pima.tr
Theoretical Quantile
Observed Quantile
20
25
30
35
40
45

$N(n$

$_1)$

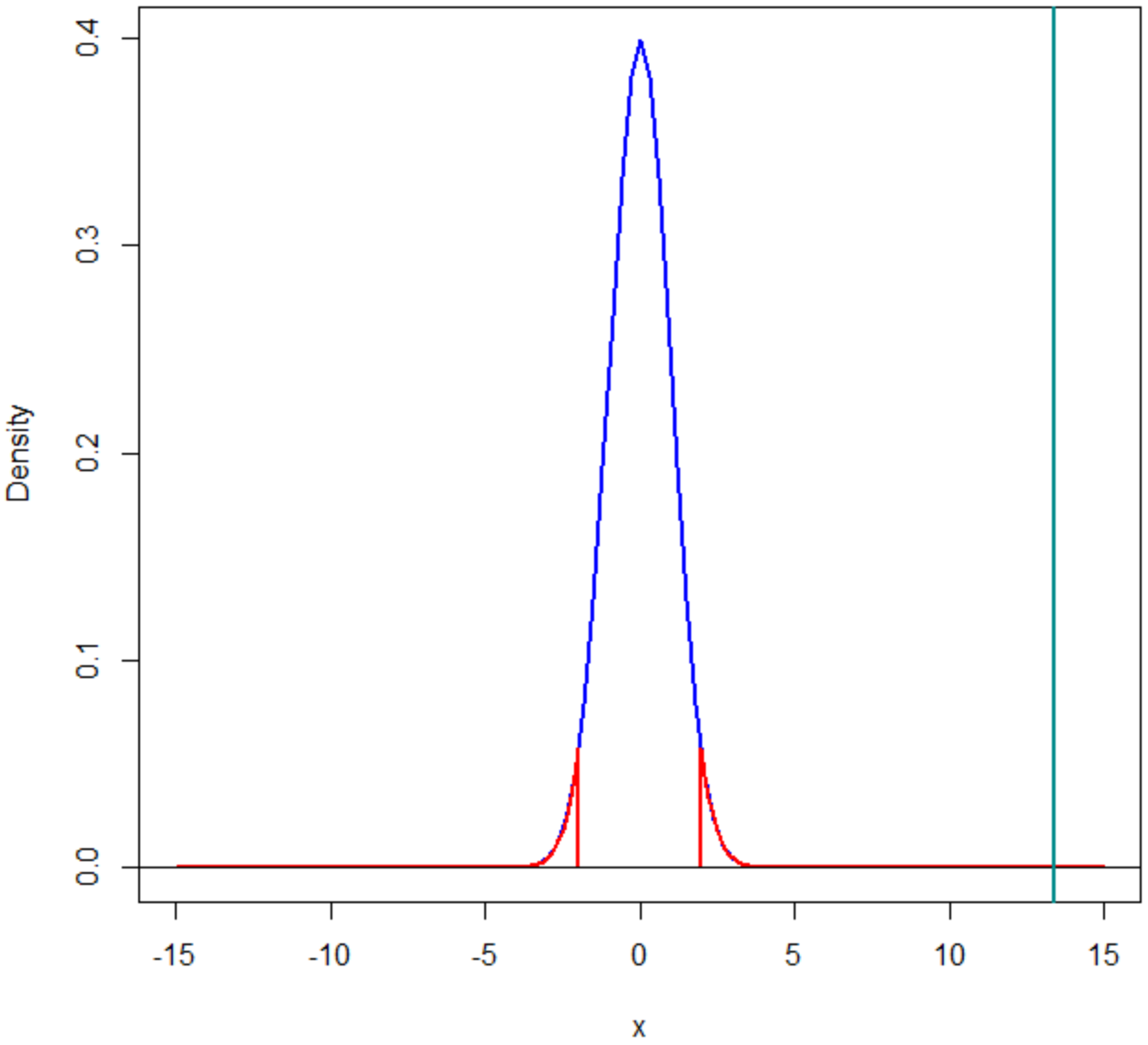
Steps in hypothesis testing: an example

- Step 3. Choose a significance level α=0.05
- Step 4. Determine the critical value
  - From n = 200, $P(|T| > C_{cri,0.05}) = 0.05$
  - Get $C_{cri,0.05} = 1.971957$
- Step 5. Calculate the test statistic

$$t = (\bar{x}_n - \mu_0) \Big/ \frac{s_n}{\sqrt{n}} = (32.31 - 26.5) \Big/ \frac{6.13}{\sqrt{100}} = 13.40$$

- Step 6. Compare the test statistic to the critical value
  - Since $|t| > C_{cri,0.05}$, we reject the null hypothesis

- Ste
- Ste
  - –
  - –
- Ste
- Ste
  crit
  - –



sis

P-value

- Often desired to see how extreme your observed data is if the null is true
  - P-value
- P-value
  - the probability that you will observe more extreme data under the null
  - The smallest significance level that your null would be rejected
- In the previous example,

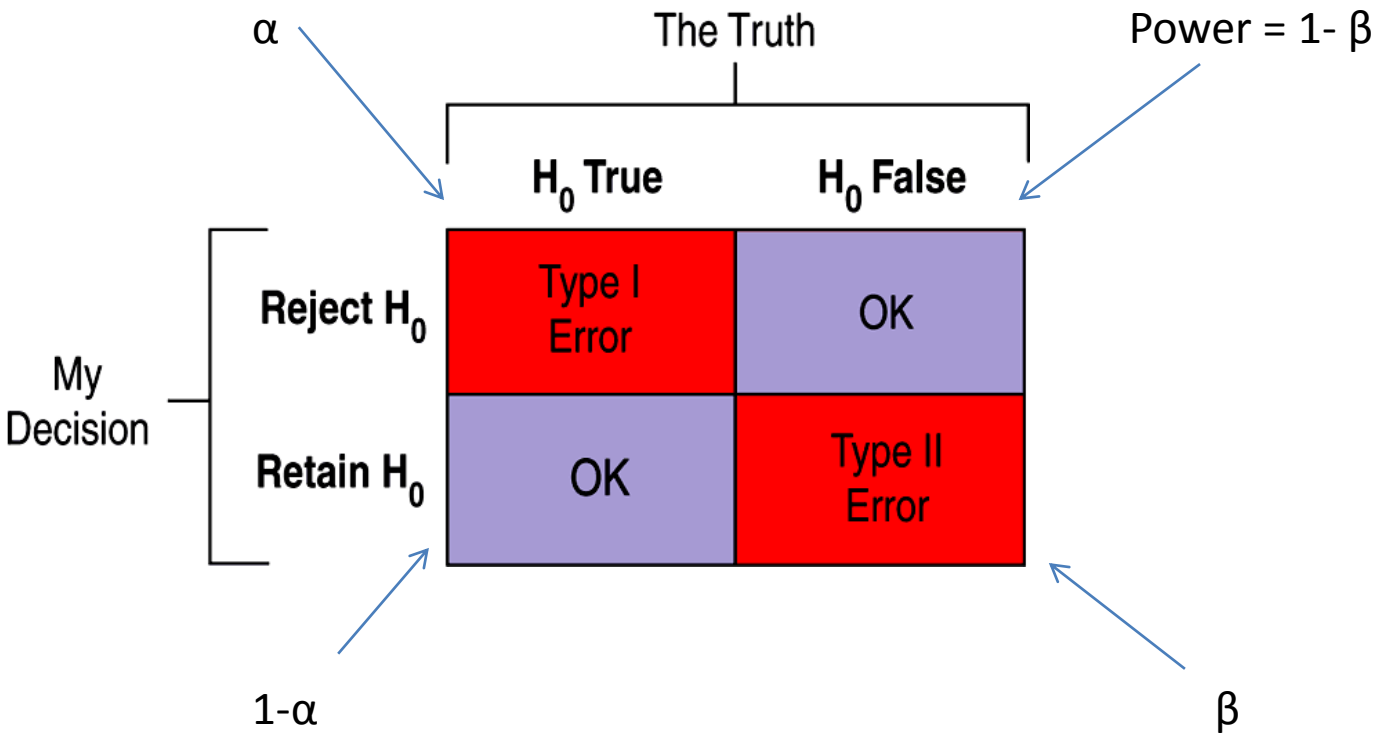
  P-value = P(|T|>t) = 1.3e-29

Making errors

- Type I error (false positive)
  - Reject the null hypothesis when the null hypothesis is true
  - The probability of Type I error is controlled by the significance level α
- Type II error (false negative)
  - Fail to reject the null hypothesis when the null hypothesis is false
  - Power = 1- probability of Type II error = 1- β
  - Power = P(reject H_0 | H_0 is false)
- Which error is more serious?
  - Depends on the context
  - In the classic hypothesis testing framework, Type I error is more serious

Making Errors

- Here's an illustration of the four situations in a hypothesis test:

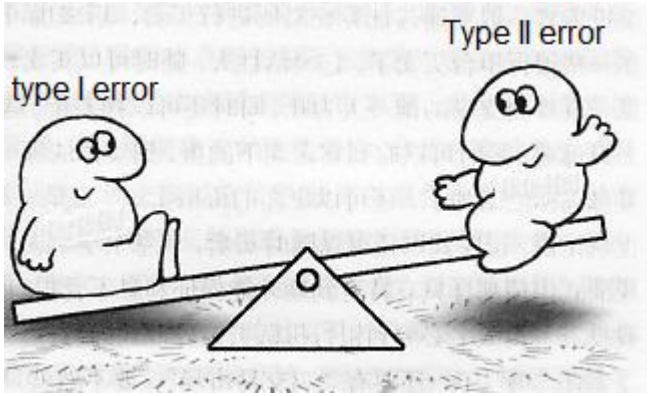


Making Errors (cont.)

- When H_0 is false and we fail to reject it, we have made a Type II error.
  - We assign the letter β to the probability of this mistake.
  - It's harder to assess the value of β because we don't know what the value of the parameter really is.
  - There is no single value for β--we can think of a whole collection of β's, one for each incorrect parameter value.

Making Errors (cont.)

- We could reduce β for *all* alternative parameter values by increasing α.
  - This would reduce β but increase the chance of a Type I error.
  - This tension between Type I and Type II errors is inevitable.
- The only way to reduce *both* types of errors is to collect more data. Otherwise, we just wind up trading off one kind of error against the other.

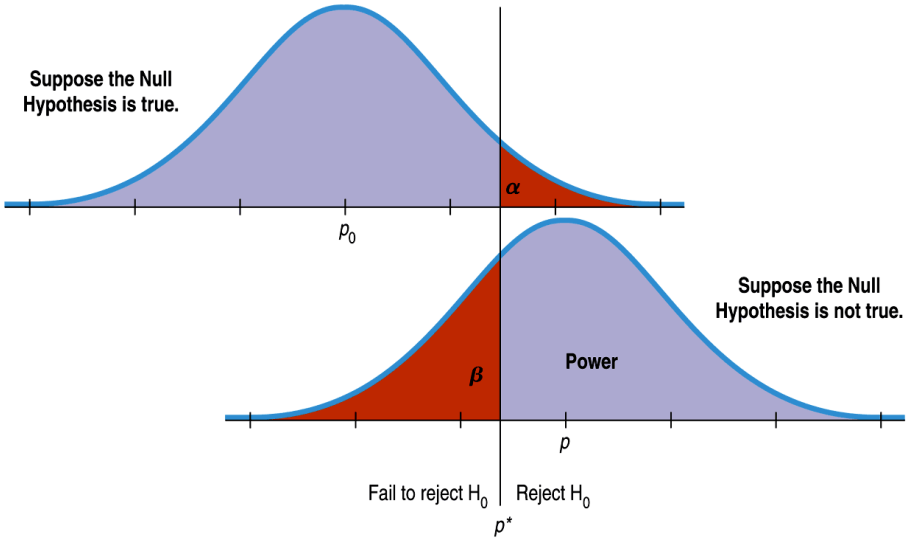


Power

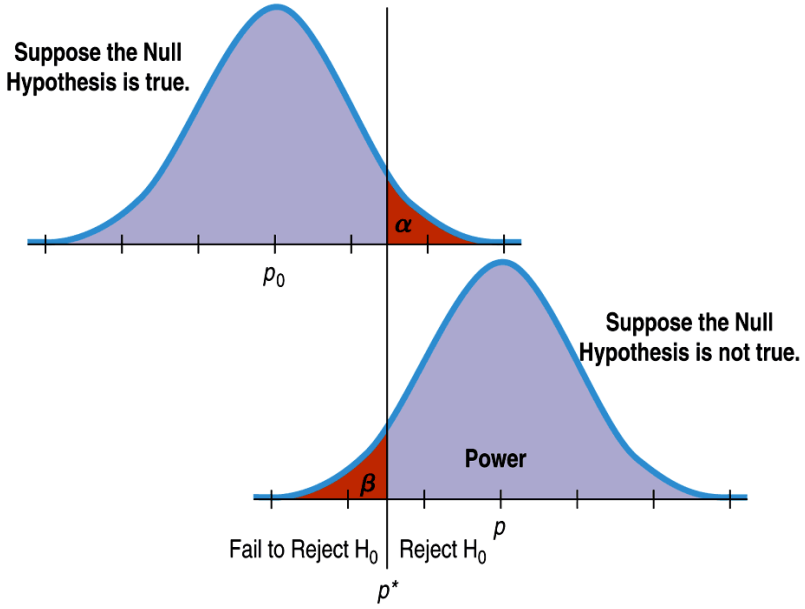
- When H_0 is false and we reject it, we have done the right thing.
  - A test's ability to detect a false hypothesis is called the power of the test.
  - The power of a test is the probability that it correctly rejects a false null hypothesis.
- When the power is high, we can be confident that we've looked hard enough at the situation.
- The power of a test is $1 - \beta$.

Reducing Both Type I and Type II Error

- Original comparison



- With a larger sample size:



Hypothesis test for single proportion

- Kantarjian et al. (2012) studied the effect of imatinib therapy on CML patients
  - CML: Chronic myelogenous leukemia (慢性粒细胞性白血病)
  - 95% of patients have ABL-BCR gene fusion
  - Imatinib was introduced to target the gene fusion
  - Since 2001, the 8-year survival rate of CML patient in chronic phase is 87%(361/415) (with Imatinib treatment)
    - Before 1990, 20%
    - 1991-2000, 45%

Hypothesis test for single proportion

- Suppose that we want to test if Imatinib can improve the 8-year survival rate
- Step 1. state the hypothesis
  - H0: $\mu=0.45$ vs H1: $\mu > 0.45$ (μ is the 8-year survival rate with Imatinib treatment)
- Step 2. Choose an appropriate test
  - Z-test based on the CLT
  - Test statistic $z = \frac{p - \mu_0}{\sqrt{\mu_0(1-\mu_0)/n}}$.
    - Follow standard normal under the null
    - Reject null if $z > C_{crt}$

Hypothesis test for single proportion

- Step 3. Choose the significance level α=0.01
- Step 4. Determine the critical value

$$P(Z > C_{cri,0.01}) = 0.05 \qquad C_{cri,0.01} = 2.33$$

- Step 5. Calculate the test statistic

$$z = (p - \mu_0) \Big/ \sqrt{\frac{\mu_0(1-\mu_0)}{n}} = 17.20$$

- Step6. Compare the test statistic with the critical value, reject the null
  - Pvalue = 1.4e-66

Comparing two populations—two sample z-test

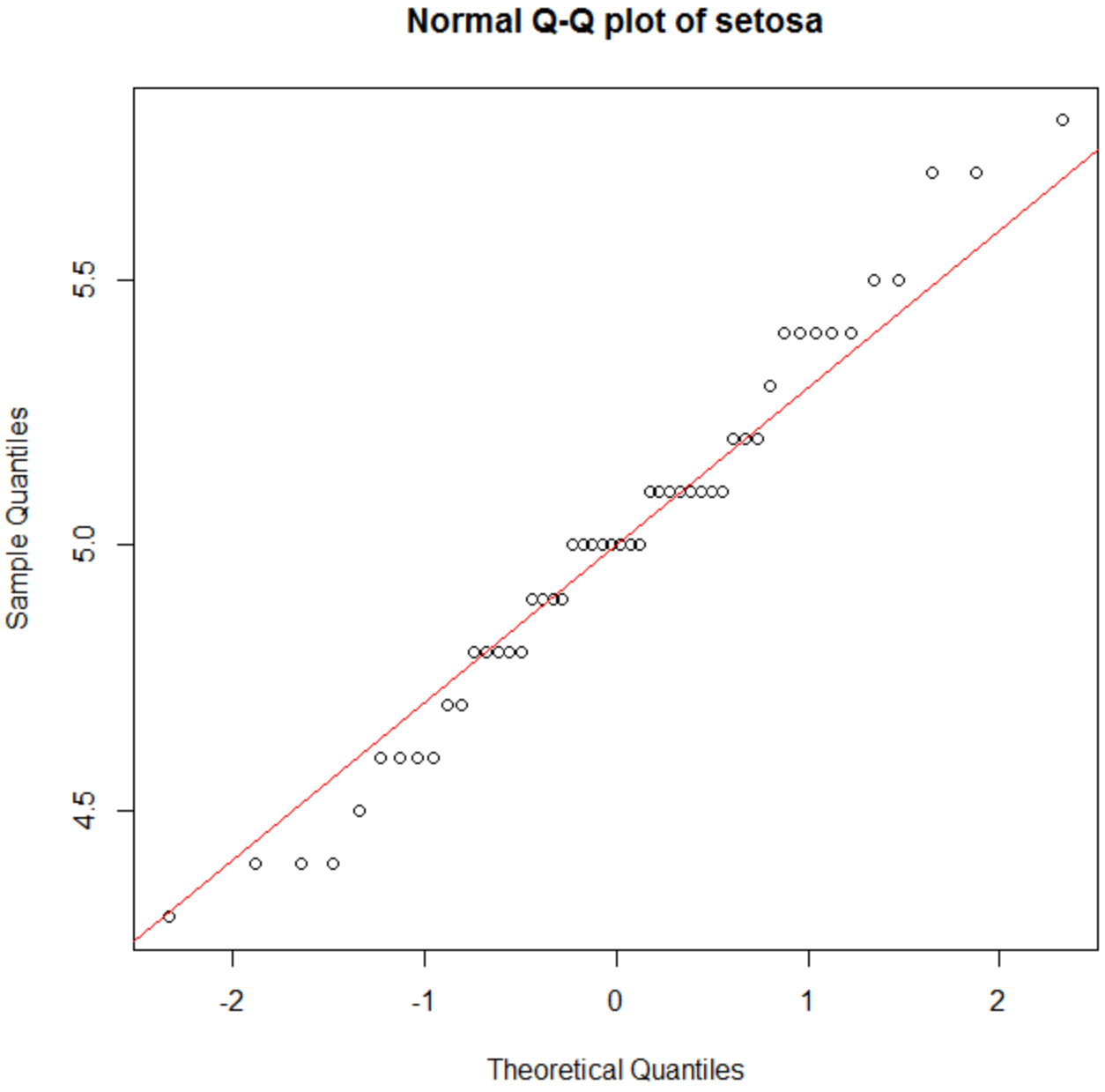
- Consider Fisher's Iris data
  - Interested to see if Sepal.Length of Setosa and versicolor are the same
  - Let μ_1 and μ_2 be their Sepal.Lengths, respectively
- State the hypothesis
  - H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 \neq \mu_2$
  - H0: $\mu_{1\,2} = \mu_1 - \mu_2 = 0$ VS H1: $\mu_{1\,2} \neq 0$

Comparing two populations—two sample z-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed with known variance ($\sigma 1 = 0.35$, $\sigma 2 = 0.38$)

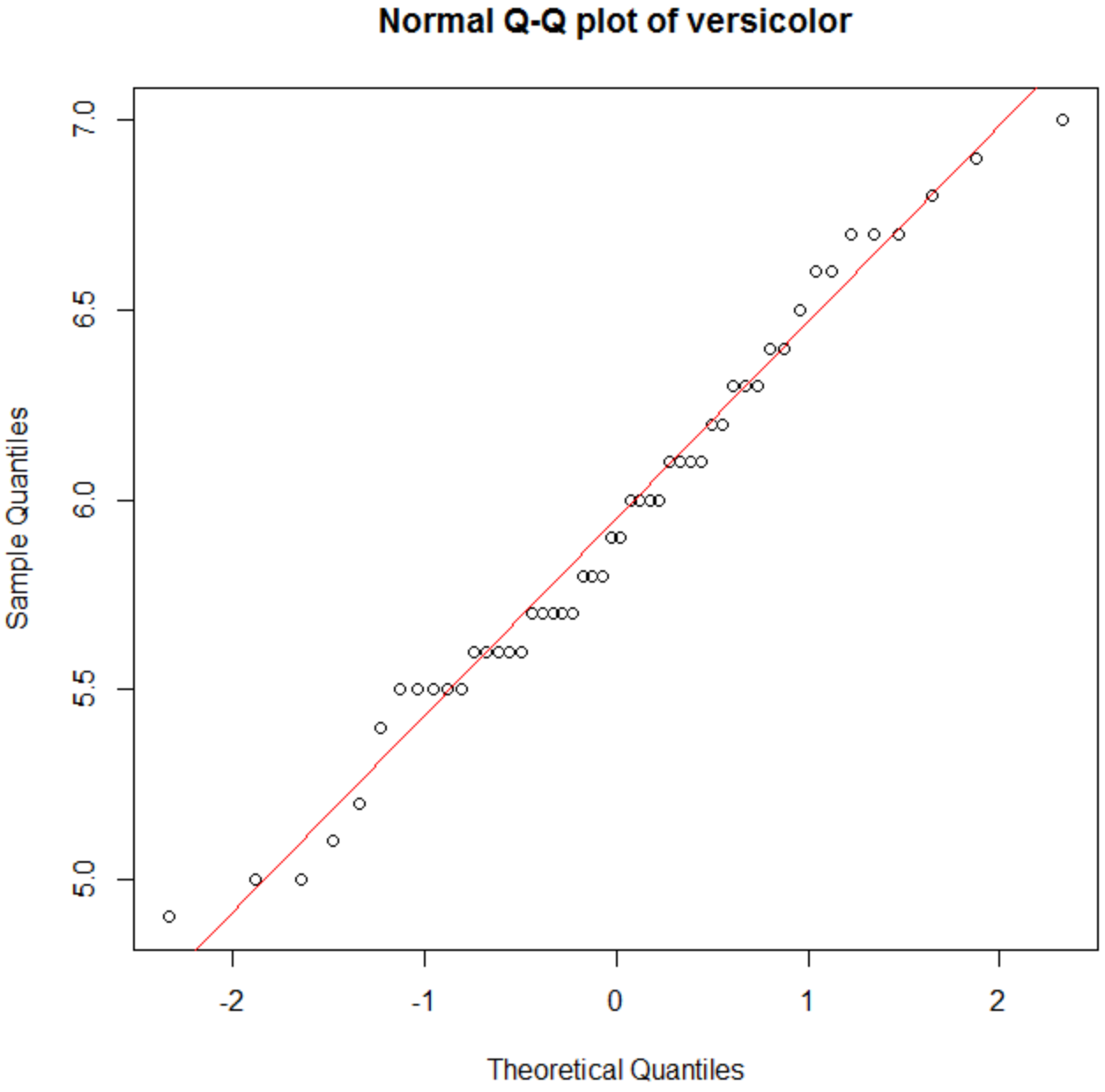
C('O

- Cho
  - F are
    r d1=
    C



C('O

- Cho
  - F are
    r d1=
    C



Comparing two populations—two sample z-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed with known variance (σ1= 0.35, σ2=0.38)
  - We have

$$\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$$
$$\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

$$Z = \frac{\bar{X}_{12} - \mu_{12}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

$$\bar{X}_{12} \sim N(\mu_1 - \mu_2,\ \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

- Significance Level α=0.01
  - $P(|Z| > C_{cri,0.01}) = 0.01$ $\qquad C_{cri,0.01} = 2.58$

Comparing two populations—two sample z-test

- Calculate the test statistics
  - z = -10.52
- |z| > 2.58, reject the NULL

$$Pvalue = P(|Z| > z) = 6.9\text{e-}26$$

- One-sided test:
  - H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 > \mu_2$
  - H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 < \mu_2$

Comparing two populations—two sample t-test

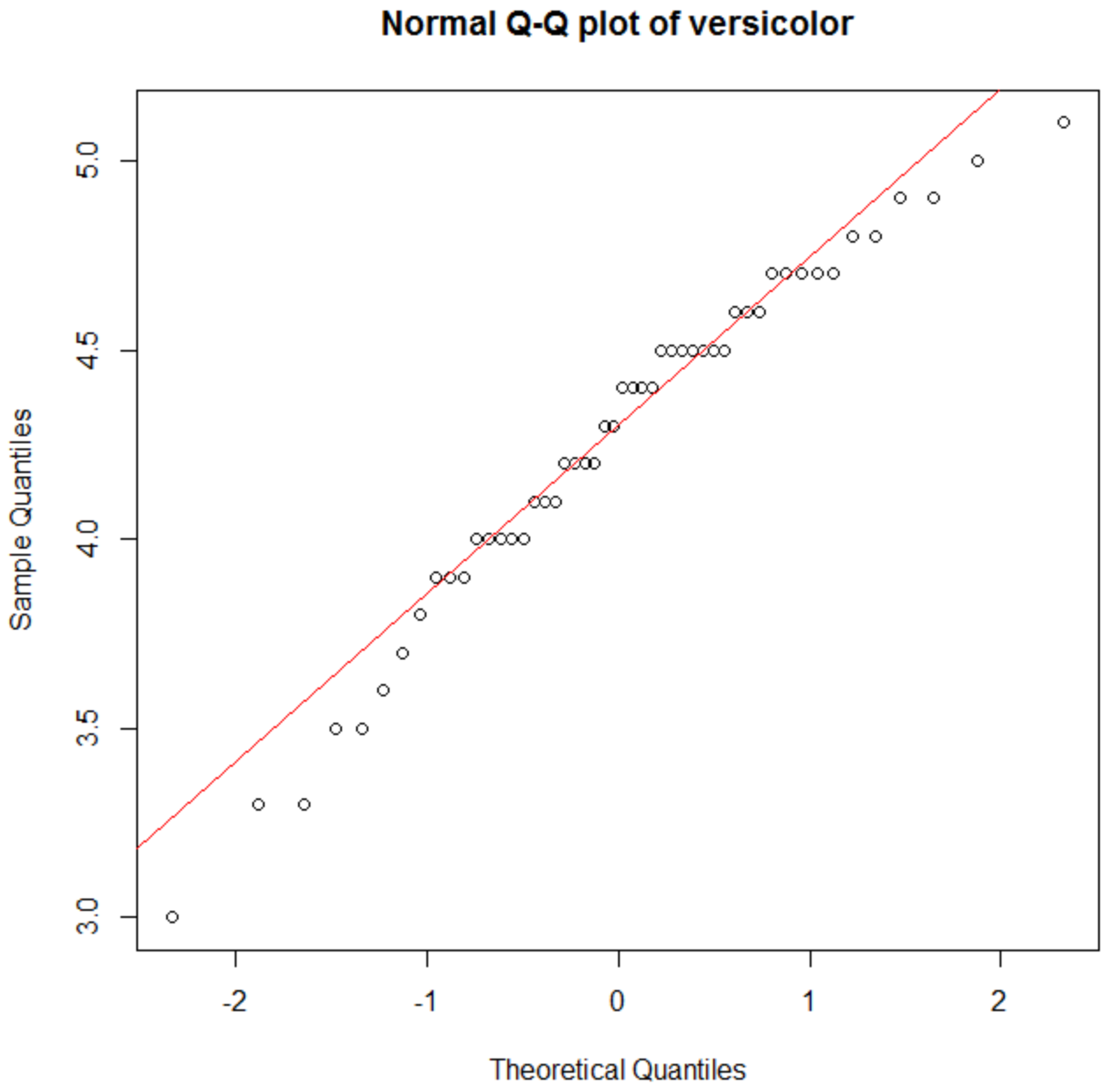
- Consider Fisher's Iris data
  - Interested to see if Petal.Length of versicolor and virginica are the same
  - Let μ_1 and μ_2 be their Petal.Length, respectively
- State the hypothesis
  - H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 \neq \mu_2$
  - H0: $\mu_{1\,2} = \mu_1 - \mu_2 = 0$ VS H1: $\mu_{1\,2} \neq 0$

Comparing two populations—two sample t-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed with unknown but equal variance

C o

- Cho
  - F are
    r l
    v



Comparing two populations—two sample t-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed with unknown but equal variance

Comparing two populations—two sample t-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed with unknown but equal variance
  - F-test for equal variance gives p-value 0.26

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ and } S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$$

$$F = \frac{S_X^2}{S_Y^2}$$

  - $t(X) = \dfrac{(\bar{X}_2 - \bar{X}_1)/\sqrt{n_1^{-1} + n_2^{-1}}}{\sqrt{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}}$ has $t_{n1+n2-2}$

Comparing two populations—two sample t-test

- Significance level 0.01
  - $P(|T| > C_{cri,0.01}) = 0.01 \qquad C_{cri,0.01} = 2.63$
- Calculate the test statistic
  - t = -12.6
- Reject the Null (|t| > 2.63)
  - $Pvalue = P(|t(X)| > t) = 3.2\text{e-}22$

Comparing two populations—two sample t-test (unequal variance)

- Consider Fisher's Iris data
  - Interested to see if Sepal.Length of Setosa and versicolor are the same
  - Let μ_1 and μ_2 be their Petal.Length, respectively
- State the hypothesis
  - H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 \neq \mu_2$
  - H0: $\mu_{1\,2} = \mu_1 - \mu_2 = 0$ VS H1: $\mu_{1\,2} \neq 0$

Comparing two populations—two sample t-test

- Choose the appropriate test
  - First Assume that the data from both groups are normally distributed
  - F-test of equal variance gives pvalue=0.009 (s1=0.35,s2=0.51)
  - Test statistic

$$t(X) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_{df}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{1}{n_1-1}(s_1^2/n_1)^2 + \frac{1}{n_2-1}(s_2^2/n_2)^2}.$$

This distribution is NOT exact

Comparing two populations—two sample t-test

- Significance level 0.01
  - $P(|T| > C_{cri,0.01}) = 0.01 \qquad C_{cri,0.01} = 2.68$
- Calculate the test statistic
  - t = -10.5
- Reject the Null (|t| > 2.68)
  - $Pvalue = P(|t(X)| > t) = 3.75\text{e-}17$

Biostatistics

BOOTSTRAPPING

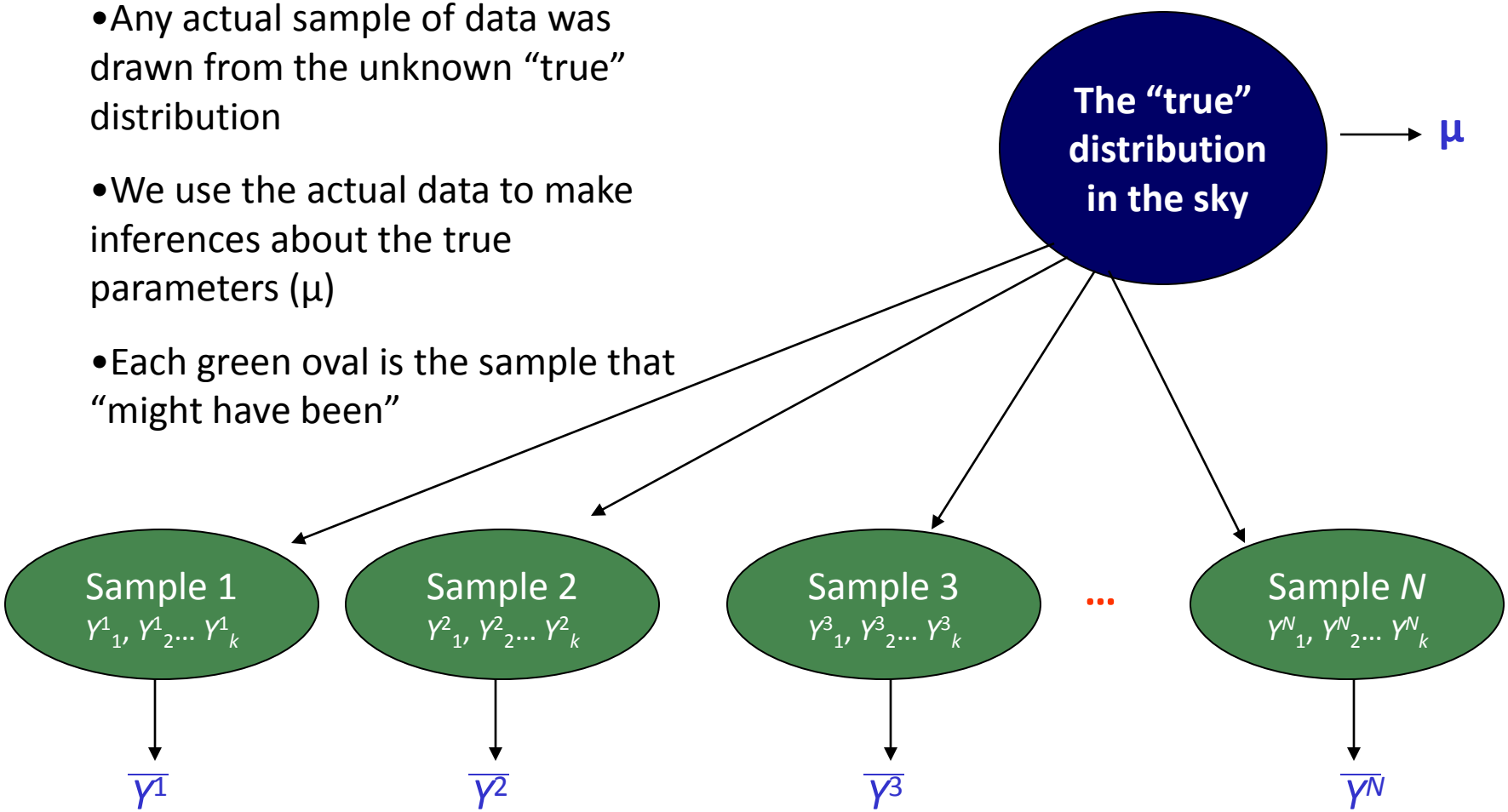
Bootstrapping

- Bootstrapping is a computational procedure for:
  - Calculating standard errors
  - Forming confidence intervals
  - Performing hypothesis tests
  - Improving predictors
- Originally proposed by Efron in 1979

The Basic Idea

Theoretical Picture

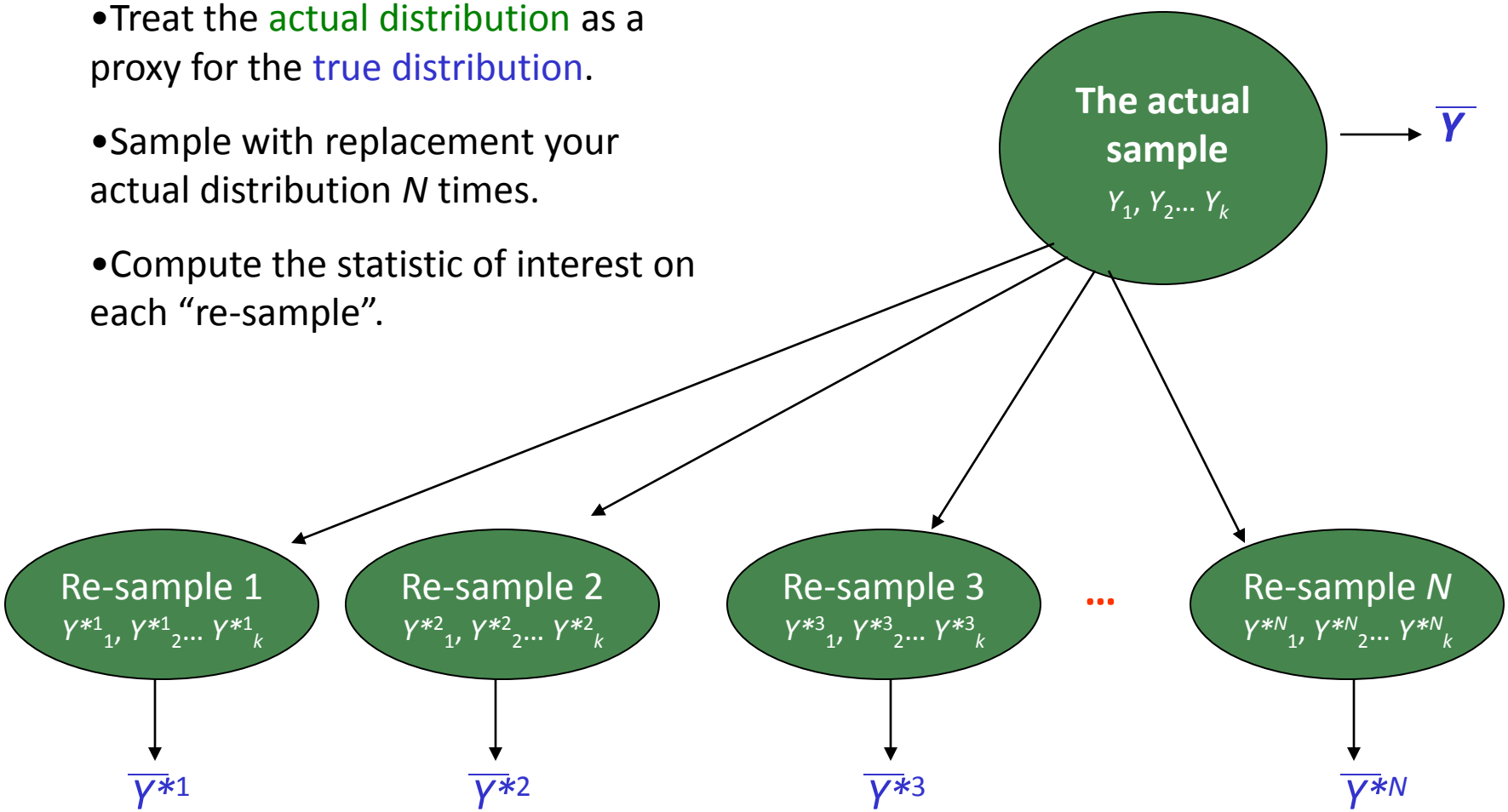
- Any actual sample of data was drawn from the unknown "true" distribution
- We use the actual data to make inferences about the true parameters (μ)
- Each green oval is the sample that "might have been"



- The distribution of our estimator ($\overline{Y}$) depends on both the true distribution *and* the size (k) of our sample

The Basic Idea

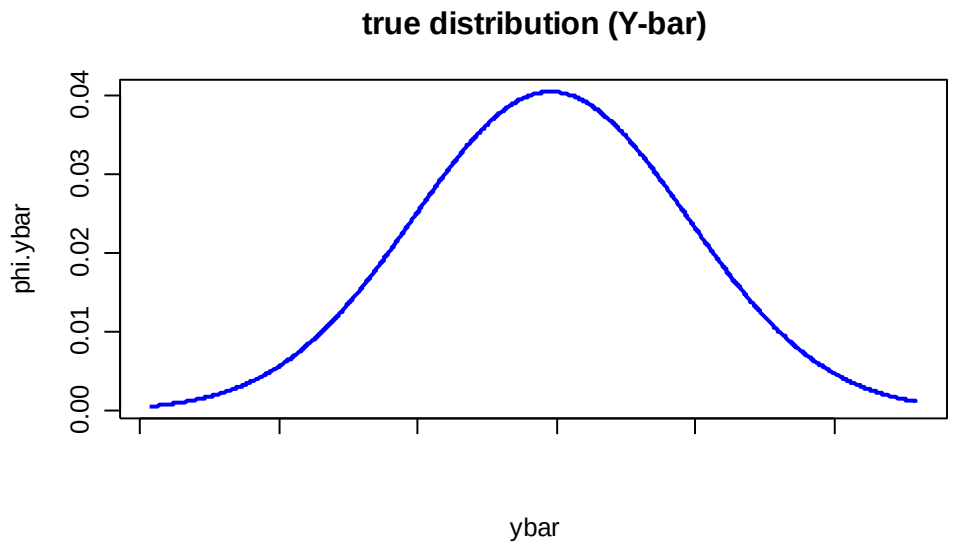
- Treat the actual distribution as a proxy for the true distribution.
- Sample with replacement your actual distribution *N* times.
- Compute the statistic of interest on each "re-sample".

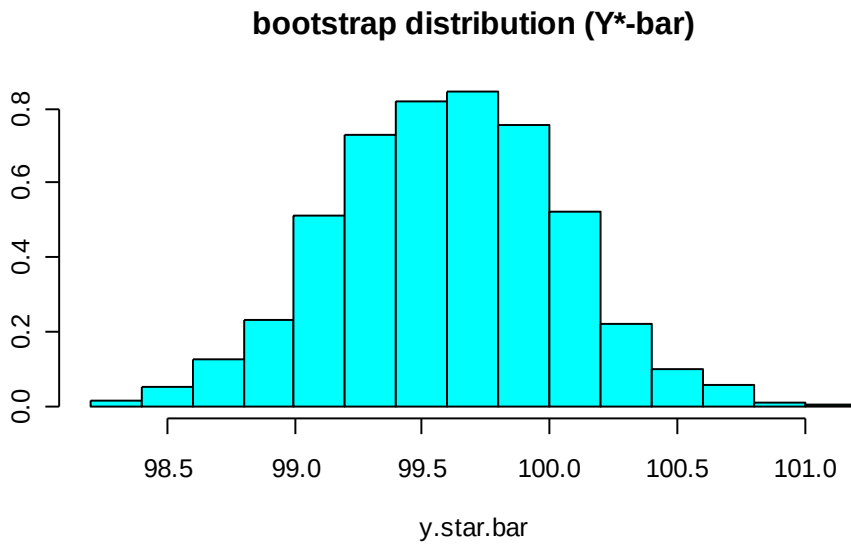


- $\{\overline{Y^*}\}$ constitutes an estimate of the *distribution* of $\overline{Y}$.

Theoretical vs. Empirical

- Graph on left: Y-bar calculated from an ∞ number of samples from the "true distribution".
- Graph on right: {Y*-bar} calculated in each of 1000 re-samples from the *empirical* distribution.
- Analogy: $\mu : \bar{Y} :: \bar{Y} : \bar{Y}^*$





Summary

- The empirical distribution – your data – serves as a proxy to the "true" distribution.
- "Resampling" means (repeatedly) sampling with replacement.
- Resampling the data is analogous to the process of drawing the data from the "true distribution".
- We can resample multiple times
    - Compute the statistic of interest *T* on each re-sample
    - We get an estimate of the distribution of *T*.

Motivating Example

- Let's look at a simple case where we all know the answer in advance.
- Pull 500 draws from the *n*(5000,100) dist.
- The sample mean ≈ 5000
  - Is a point estimate of the "true" mean μ.
  - But how sure are we of this estimate?
- From theory, we know that:

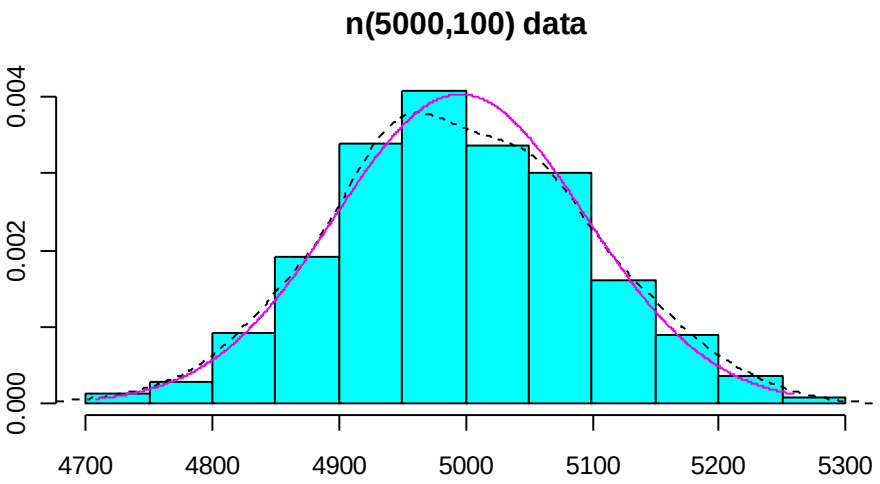
$$s.d.(\overline{X}) = \sigma / \sqrt{N} \approx 100 / \sqrt{500} \approx 4.47$$

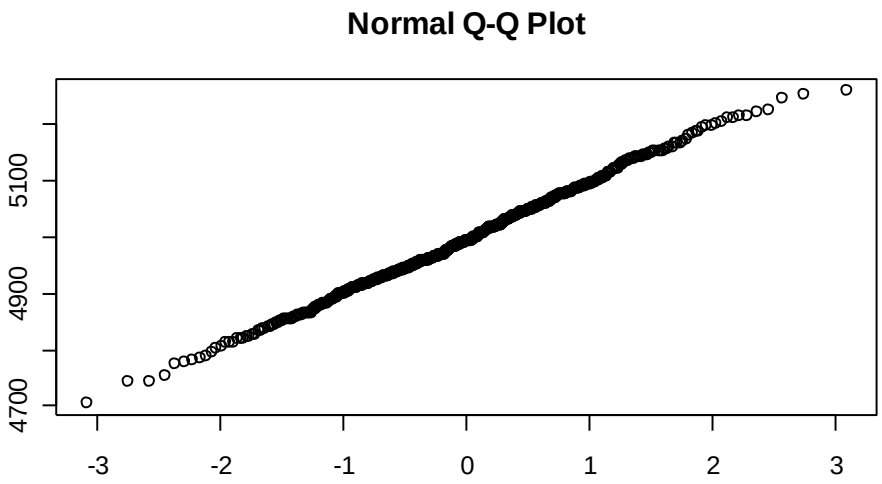
raw data	
statistic	**value**
#obs	500
mean	**4995.79**
sd	**98.78**
2.5%*ile*	4812.30
97.5%*ile*	5195.58

Visualizing the Raw Data

- 500 draws from *n*(5000,100)
- Look at summary statistics, histogram, probability density estimate, QQ-plot.
- … looks pretty normal

raw data	
statistic	**value**
#obs	500
mean	**4995.79**
sd	**98.78**
2.5%*ile*	4812.30
97.5%*ile*	5195.58





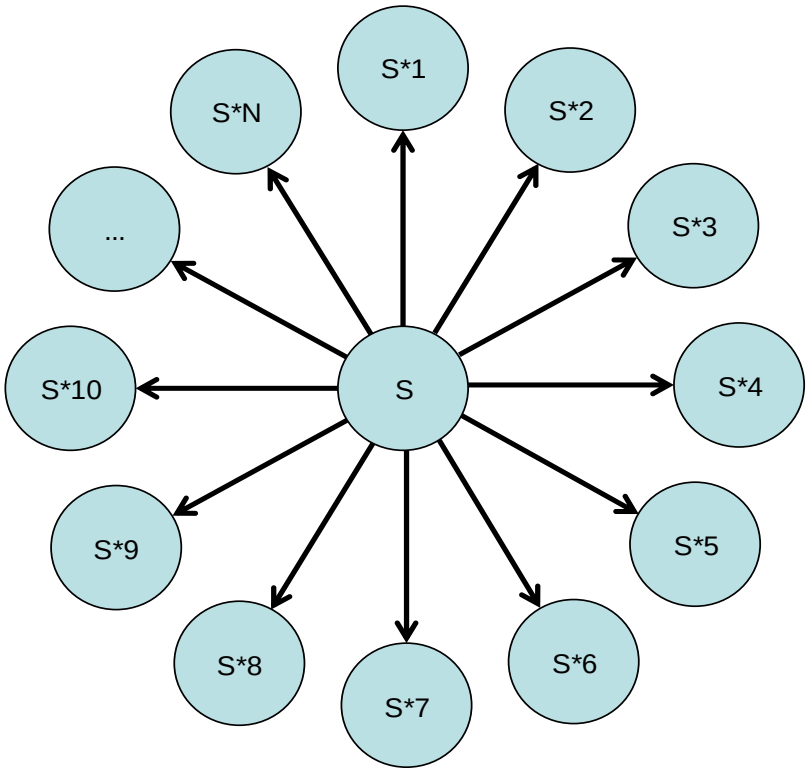
Sampling With Replacement

Now let's use resampling to estimate the s.d. of the sample mean (≈4.47)

- Draw a data point at random from the data set.
  - Then throw it back in
- Draw a second data point.
  - Then throw *it* back in…
- Keep going until we've got 500 data points.
  - You might call this a "pseudo" data set.
- This is not merely re-sorting the data.
  - Some of the original data points will appear more than once; others won't appear at all.

Resampling

- Sample with replacement 500 data points from the original dataset S
  - Call this S^*_1
- Now do this 999 more times!
  - $S^*_1, S^*_2, \ldots, S^*_{1000}$
- Compute X-bar on each of these 1000 samples.



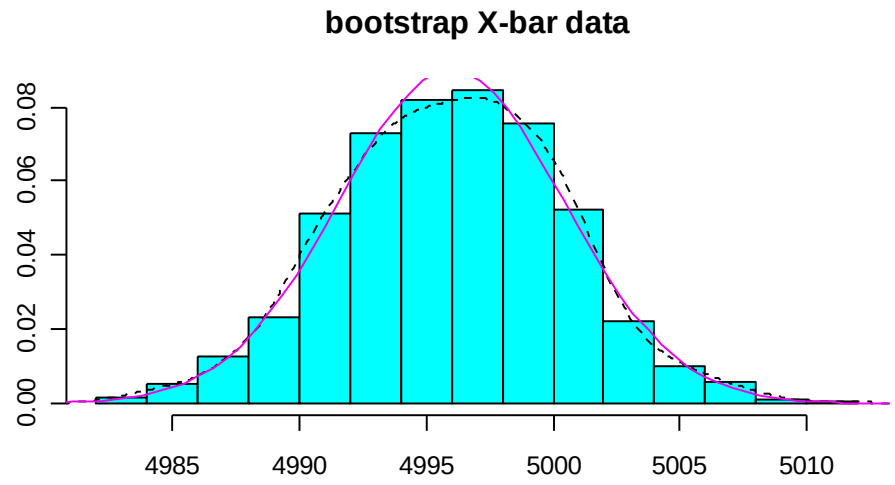
R Code

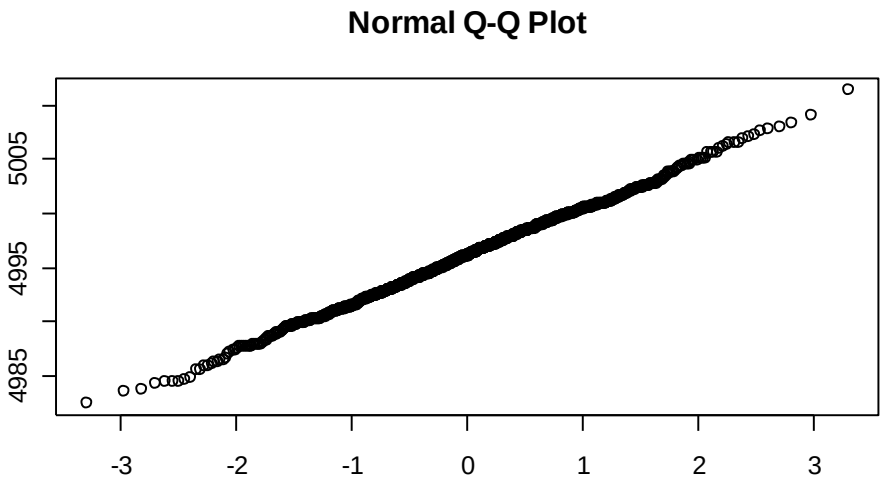
```
norm.data <- rnorm(500, mean=5000, sd=100)
boots <- function(data, R){
b.avg <<- c(); b.sd <<- c()
for(b in 1:R) {
        ystar <- sample(data,length(data),replace=T)
        b.avg <<- c(b.avg,mean(ystar))
        b.sd  <<- c(b.sd,sd(ystar))}
}
boots(norm.data, 1000)
```

Results

- From theory we know that **X-bar ~ *n*(5000, 4.47)**
- Bootstrapping estimates this pretty well!
- And we get an estimate of the *whole distribution*, not just a confidence interval.

raw data		X-bar	
statistic	value	theory	bootstrap
#obs	500	1,000	1,000
mean	**4995.79**	**5000.00**	**4995.98**
sd	**98.78**	**4.47**	**4.43**
2.5%*ile*	4705.08	4991.23	4987.60
97.5%*ile*	5259.27	5008.77	5004.82





Two Ways of Looking at a Confidence Interval

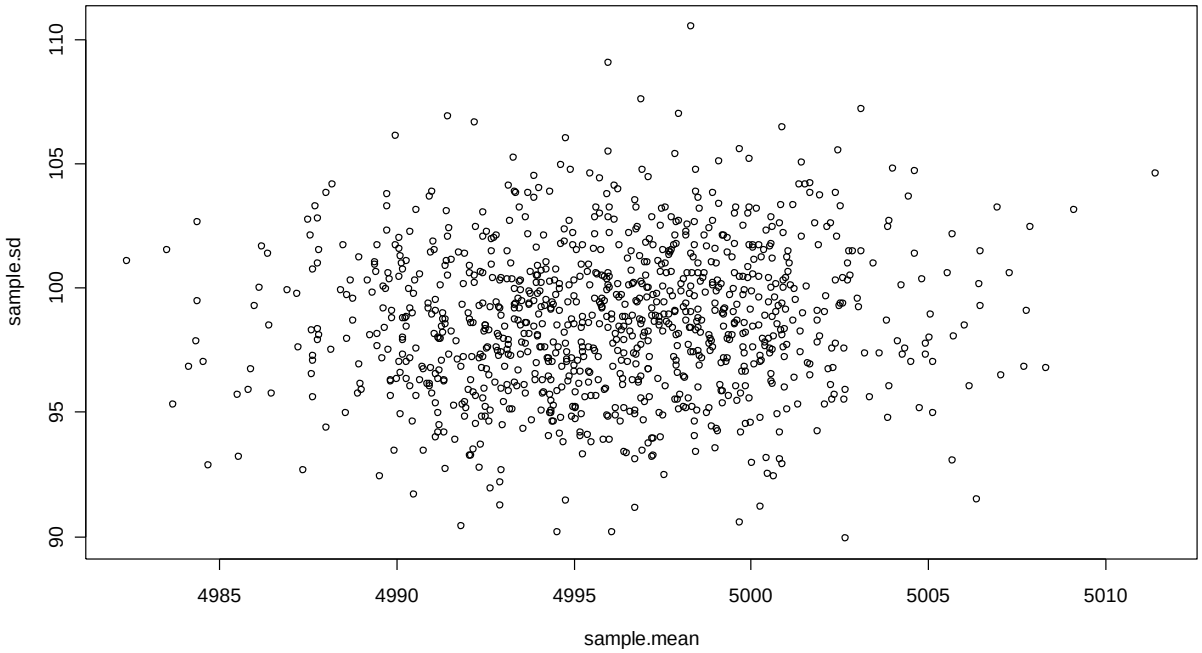
- Approximate normality assumption
  - X-bar ±2*(bootstrap dist s.d.)
- Percentile method
  - Just take the desired percentiles of the bootstrap histogram.
  - More reliable in cases of asymmetric bootstrap histograms.

```
mean(norm.data) - 2 * sd(b.avg)
[1] 4986.926
mean(norm.data) + 2 * sd(b.avg)
[1] 5004.661
```

raw data		X-bar	
statistic	value	theory	bootstrap
#obs	500	1,000	1,000
mean	**4995.79**	**5000.00**	**4995.98**
sd	**98.78**	**4.47**	**4.43**
2.5%*ile*	4705.08	4991.23	4987.60
97.5%*ile*	5259.27	5008.77	5004.82

And a Bonus

- Note that we can calculate both the mean and standard deviation of each pseudo-dataset.
- This enables us to estimate the correlation between the mean and s.d.
- Normal distribution is not skew ➔ mean, s.d. are uncorrelated.
- Our bootstrapping experiment confirms this.



More Interesting Examples

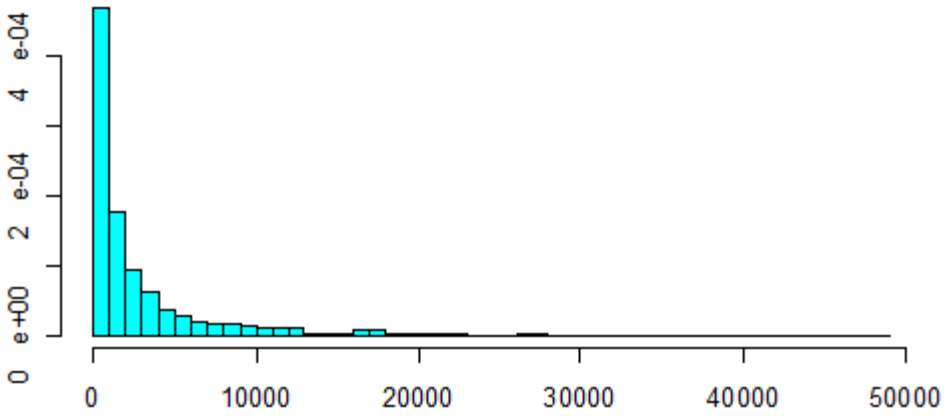
- We've seen that bootstrapping replicates a result we know to be true from theory.
- Often in the real world we either don't know the 'true' distributional properties of a random variable…
- …or are too busy to find out.
- This is when bootstrapping really comes in handy.

Skewed Data

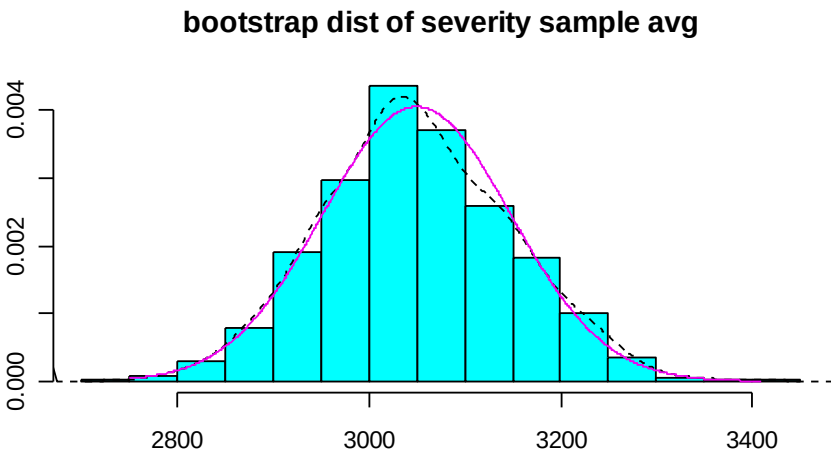
- 2700 data points.
  - **Mean = 3052, Median = 1136**

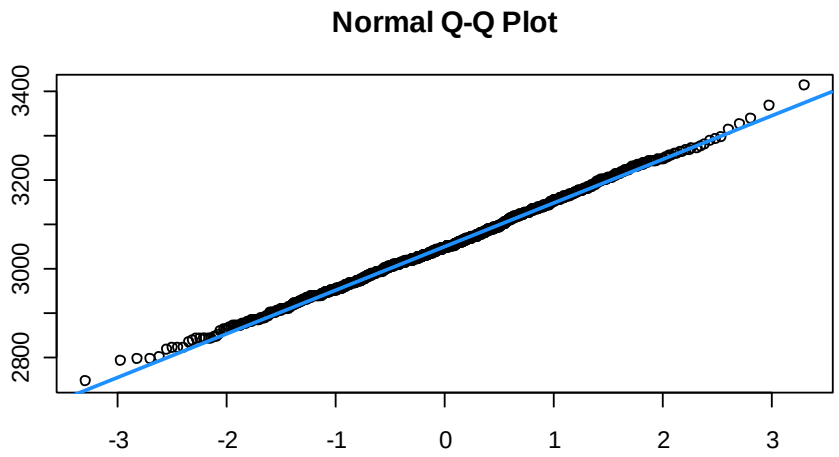
0%	25%	50%	75%	100%
51.84	482.42	1136.10	3094.09	48346.82

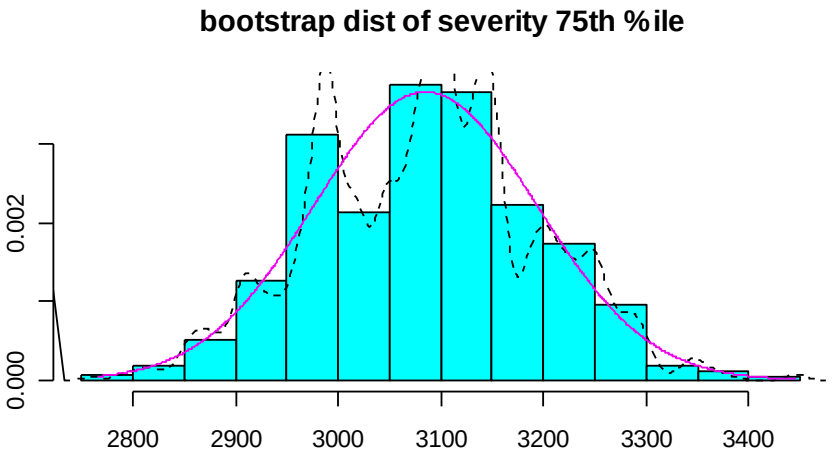
- Let's estimate the distributions of the sample mean & 75th *%ile.*
- Gamma? Lognormal? Don't need to know.

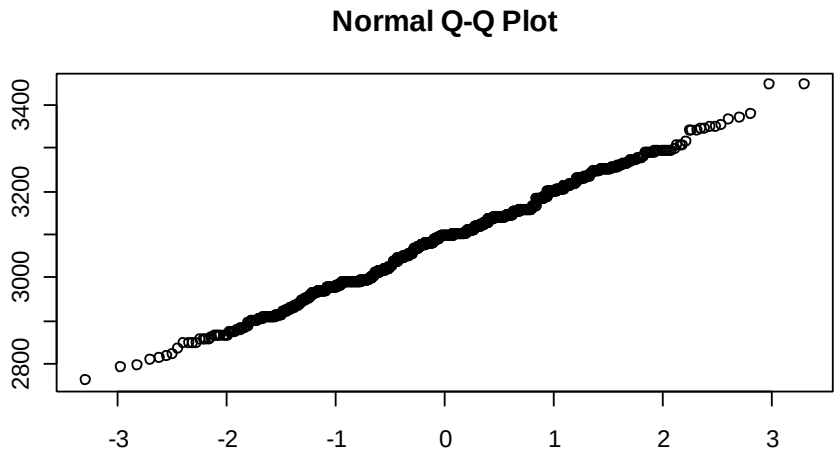


Bootstrapping Sample Avg, 75th %*ile*



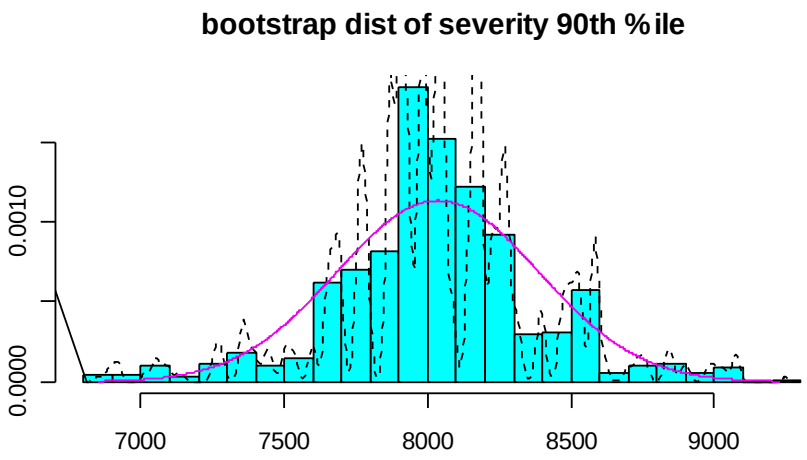


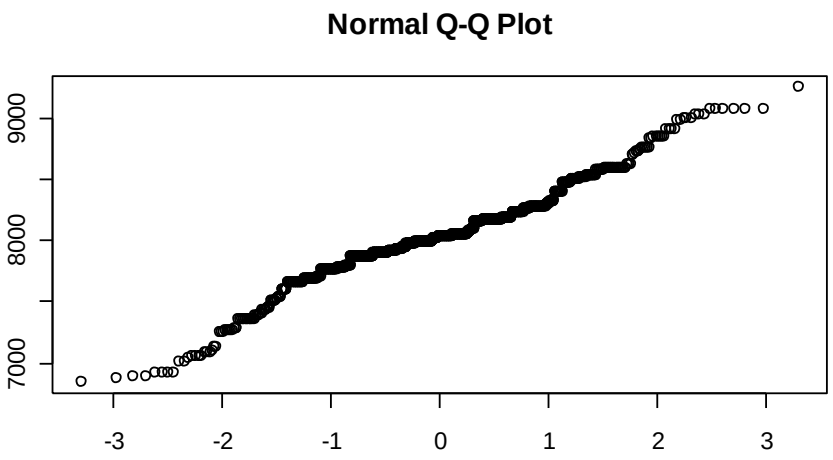




What about the 90th %*ile*?

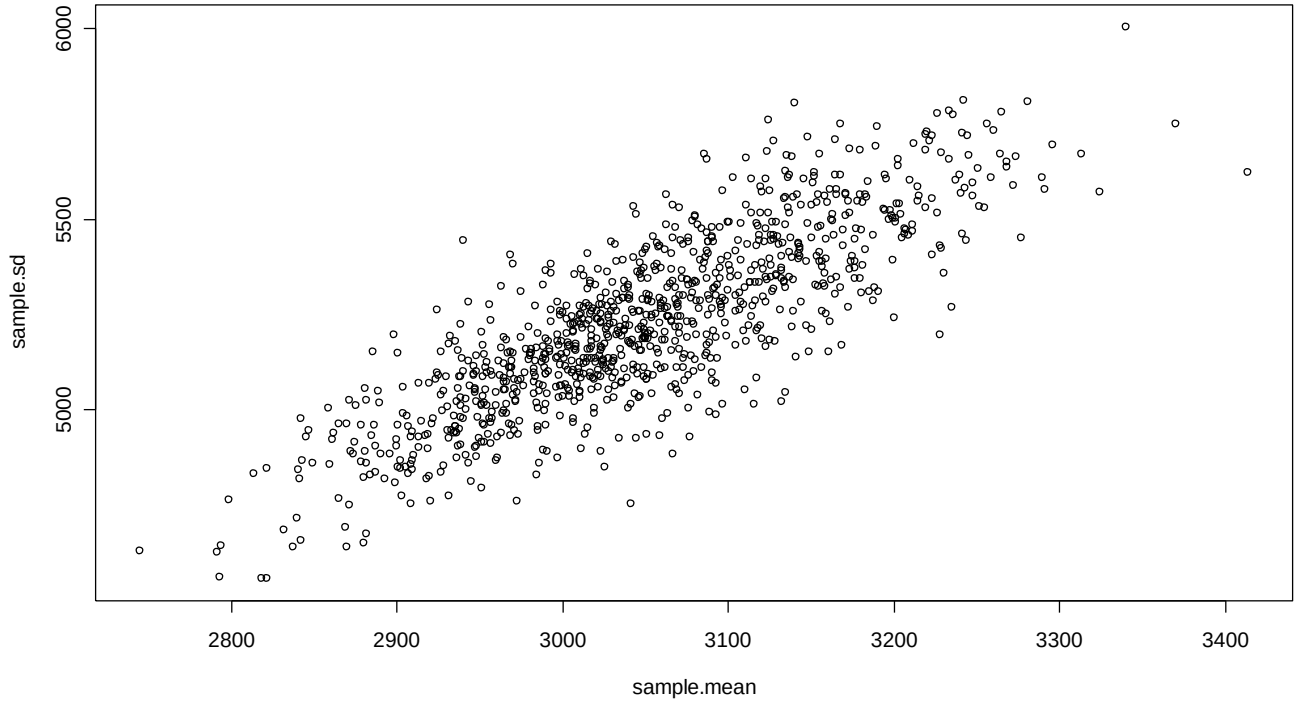
- So far so good – bootstrapping shows that many of our sample statistics – even average severity! – are approximately normally distributed.
- But this breaks down if our statistics is not a "smooth" function of the data…
  - Often in the loss reserving we want to focus our attention way out in the tail…
- 90th %*ile* is an example.





Variance Related to the Mean

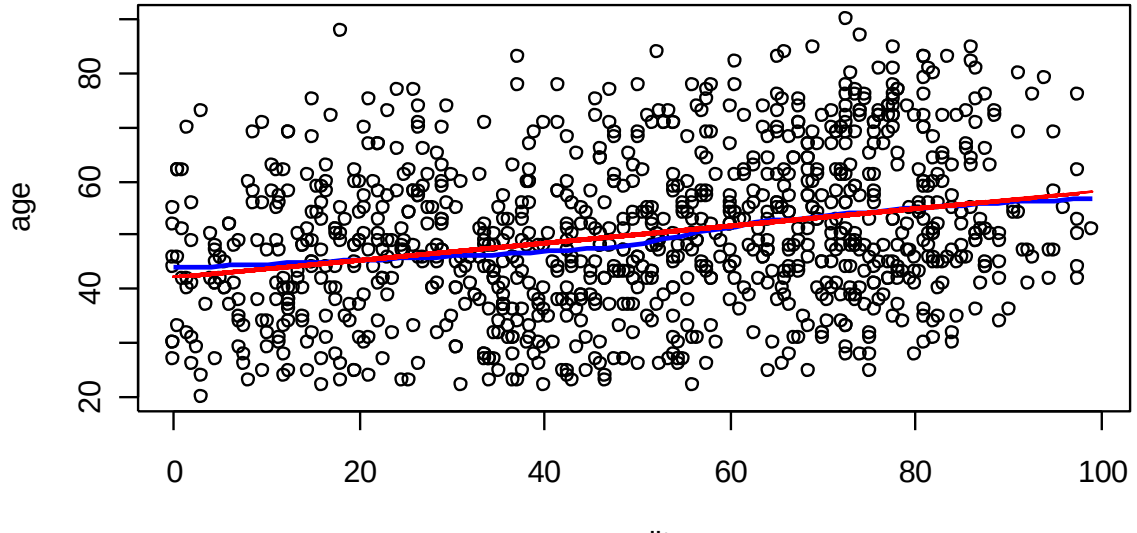
- As with the normal example, we can calculate both the sample average and s.d. on each pseudo-dataset.
- This time (as one would expect) the variance is a function of the mean.



Bootstrapping a Correlation Coefficient #1

- About 700 data points
- Credit on a scale of 1-100
  - *1 is worst; 100 is best*
- Age, credit are linearly related
  - See plot
- $R^2 \approx .08$ ➔ $\rho \approx .28$
  - Older people tend to have better credit
- What is the confidence interval around ρ?

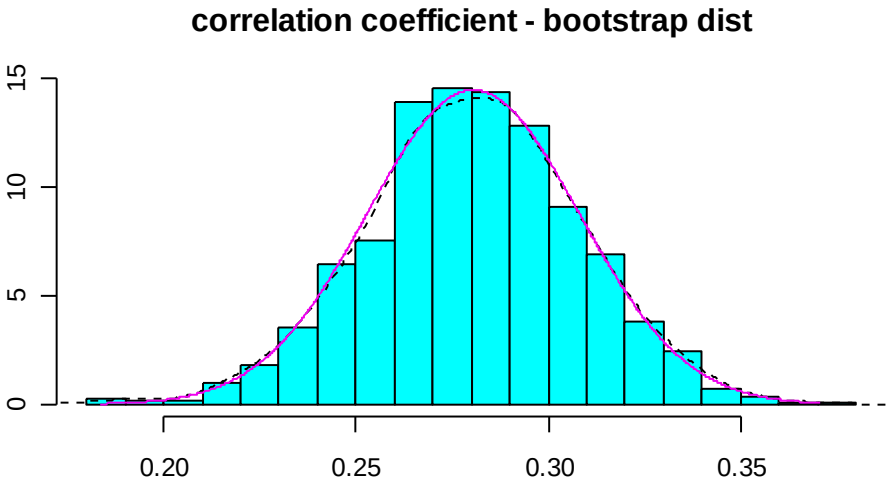
Plot of Age vs Credit

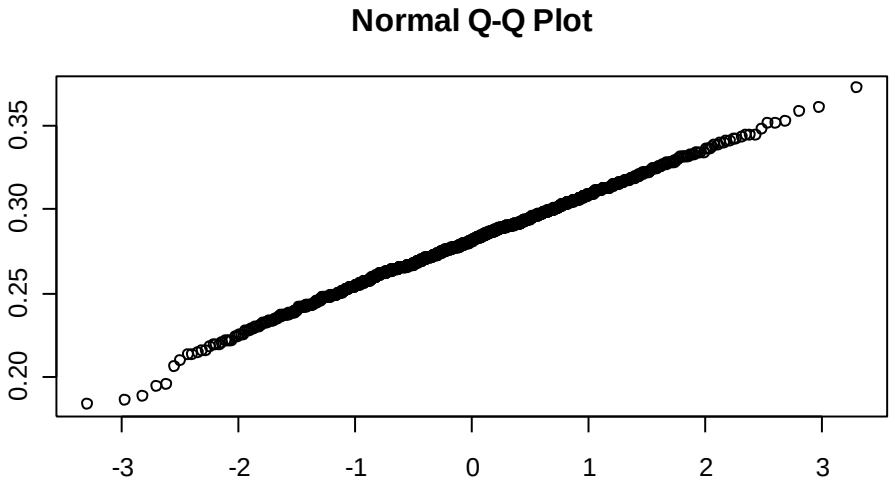


Bootstrapping a Correlation Coefficient #1

- ρ appears normally distributed.
  - ρ ≈ .28
  - s.d.(ρ) ≈ .028
- Both confidence interval calculations agree fairly well:

```
> quantile(boot.avg,probs=c(.025,.975))
    2.5%     97.5%
0.2247719 0.3334889
> rho - 2*sd(boot.avg); rho + 2*sd(boot.avg)
0.2250254   0.3354617
```

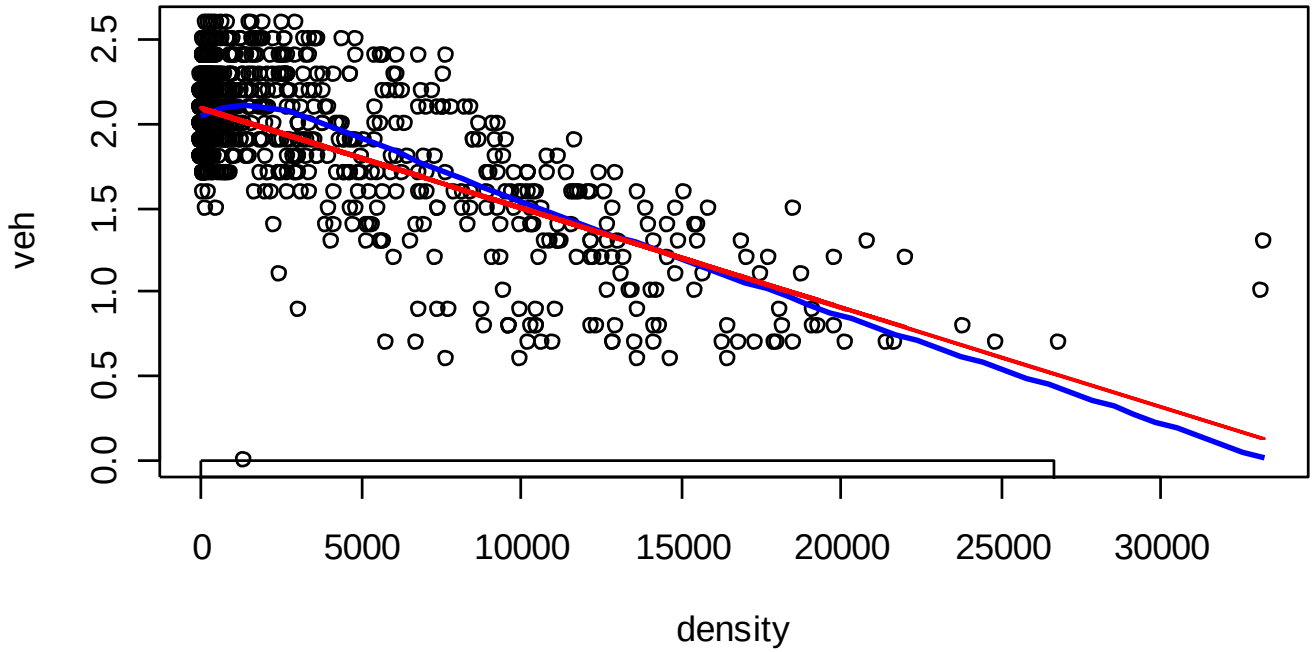




Bootstrapping a Correlation Coefficient #2

- Let's try a different example.
- ≈1300 zip-code level data points
- Variables: population density, median #vehicles/HH
  - $R^2 \approx .50$; $\rho \approx -.70$

Median #Vehicles vs Pop Density



Bootstrapping a Correlation Coefficient #2

- ρ more skew.
  - ρ ≈ -.70
  - 95% conf interval: **(-.75, -.67)**
  - Not symmetric around ρ
  - Effect becomes more pronounced the higher the value of ρ.