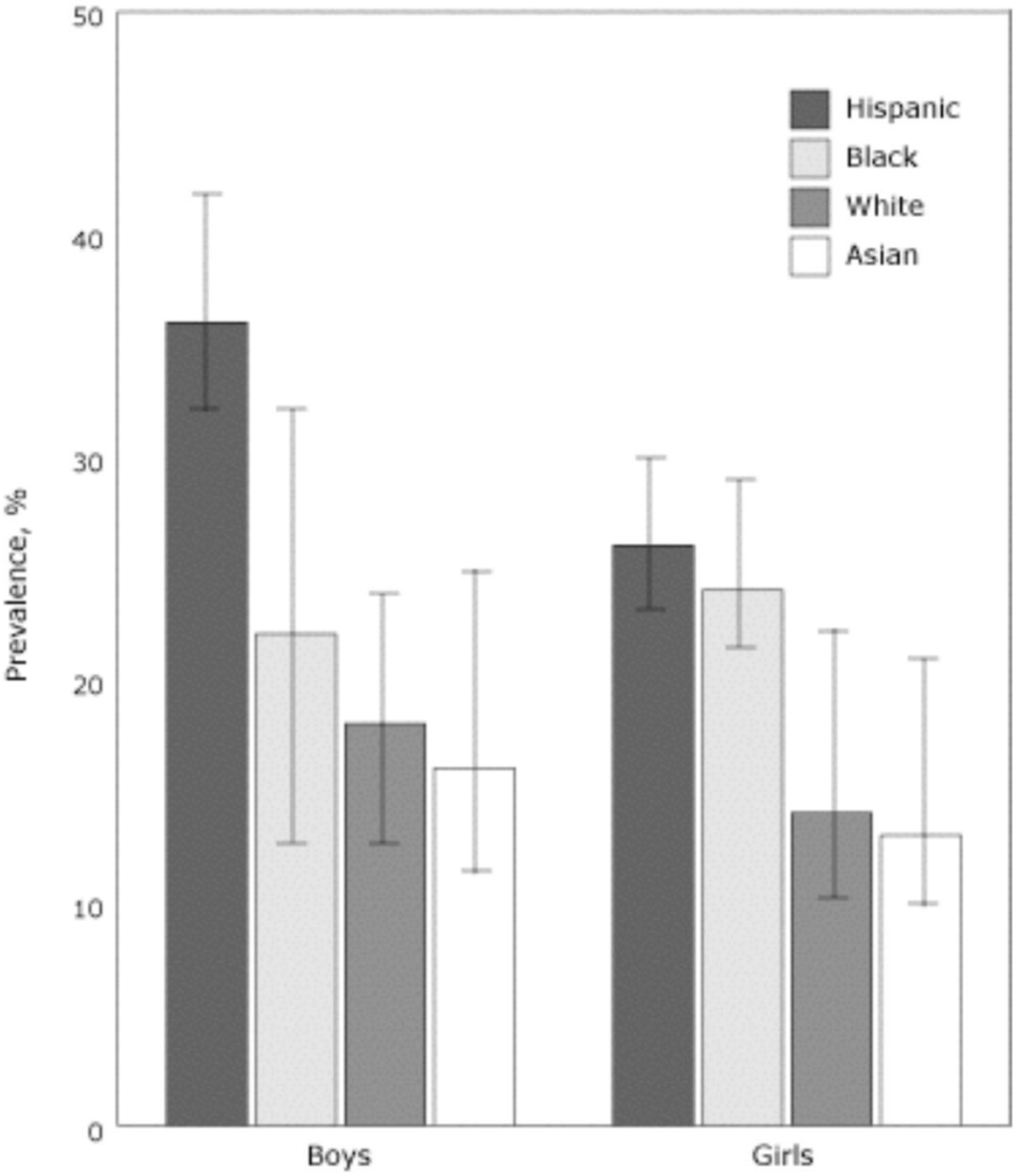
Confidence Intervals

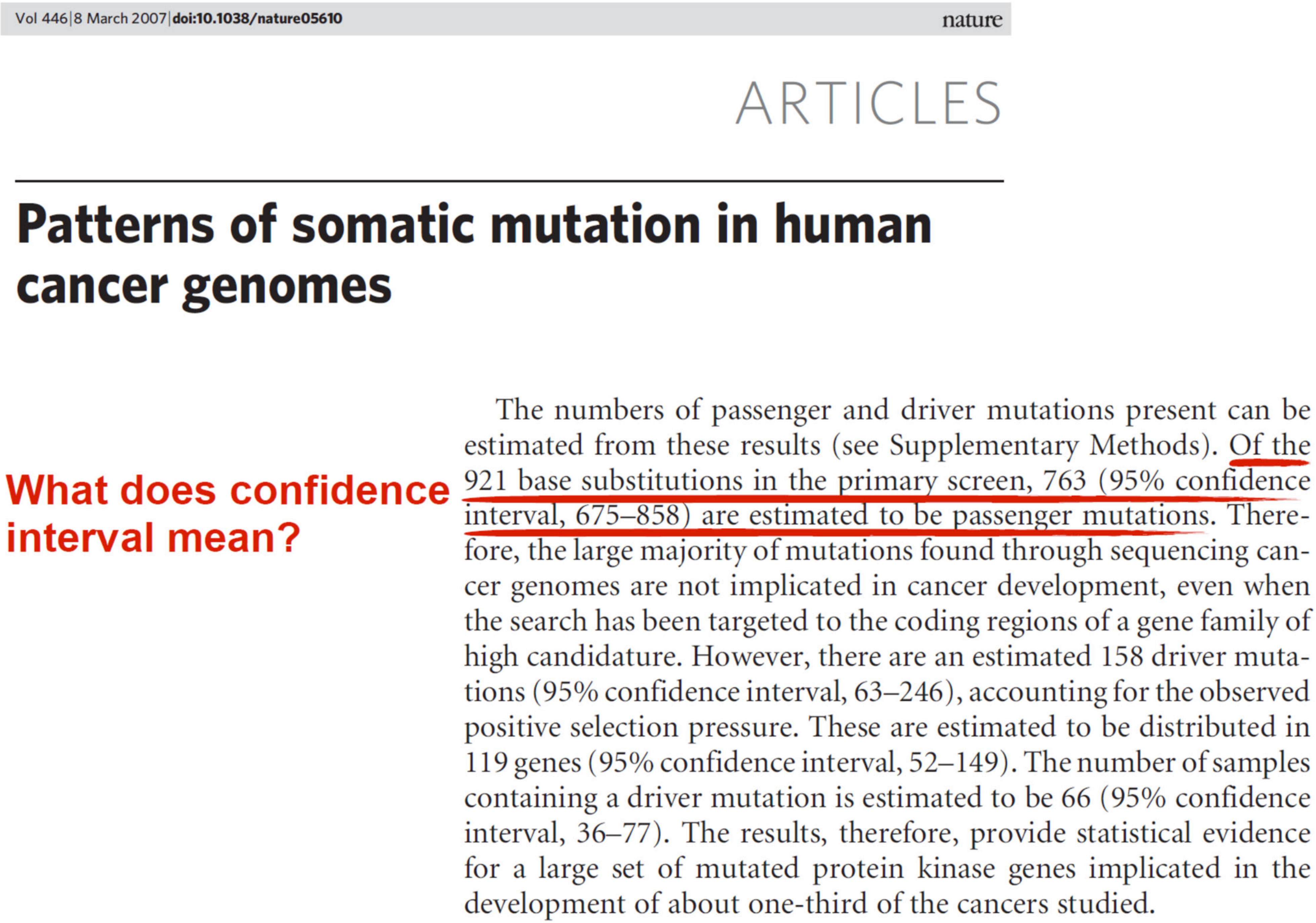


Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

What do those bars actually mean?

What does confidence interval mean?



Vol 446|8 March 2007|doi:10.1038/nature05610 nature

ARTICLES

Patterns of somatic mutation in human cancer genomes

The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.

- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

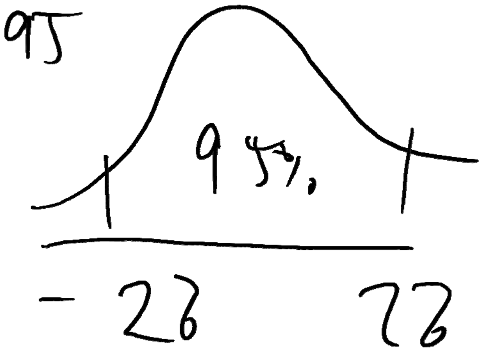
- Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate.

Two-sided confidence intervals

- Calculated based on the sample X_1, X_2,…,X_n
- Characterized by:
  - lower- and upper- confidence limits *L* and *R*
  - the confidence coefficient *1-α*
- Objective: for two-sided confidence interval, find L and R such that
  - Prob*(μ>R)=α/2*
  - Prob*(μ<L)=α/2*
  - Therefore, Prob*(L<μ<R)=1-α*
- For one-sided confidence interval, say, upper bound of *μ ,* find R that
  - Prob*(μ>R)=α*
- **Assume standard deviation sigma is known**

Consider $1 - \alpha = 95\% = 0.95$

$\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$



$z_{\alpha/2} = 1.96 \Rightarrow \text{Prob}(Z > z_{\alpha/2}) = \frac{\alpha}{2}$

$$\text{Prob}\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\text{Prob}\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

For one sided lower bound on μ

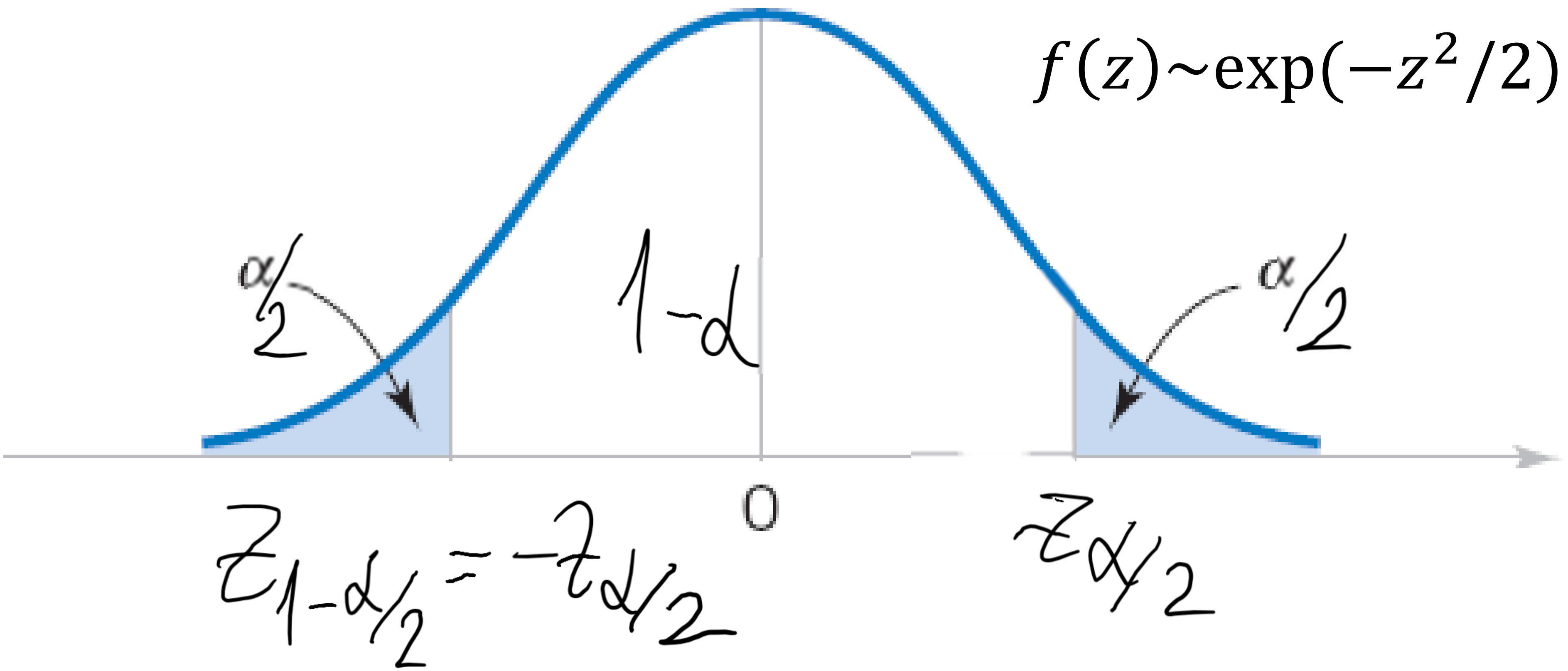
$$\text{Prob}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\underline{\underline{\underline{\alpha}}}}\right) \rightarrow$$

$$\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

$z_\alpha = 1.65 <$
$z_{\alpha/2} = 1.96$

Confidence Interval on the Population Mean, Variance Known

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



$f(z) \sim \exp(-z^2/2)$

Exercise

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface.

Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2.

Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface. Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2. Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

a) 95% CI for μ, $n = 5$ $\sigma = 0.66$ $\bar{x}$ = 3.372, $z = 1.96$

$$\bar{x} - z\sigma/\sqrt{n} \le \mu \le \bar{x} + z\sigma/\sqrt{n}$$

$$3.372 - 1.96(0.66/\sqrt{5}) \le \mu \le 3.372 + 1.96(0.66/\sqrt{5})$$

$$2.79 \le \mu \le 3.95$$

b) Width is $2z\sigma/\sqrt{n} = 0.55$, therefore n = [2zσ/0.55]2 = [2(1.96)(0.66)/0.55]2 = 22.13

Round up to $n = 23$.

Matlab exercise

- 1000 labs measured average P53 gene expression using n=20 samples drawn from the Gaussian distribution with mu=3; sigma=2;
- Each lab found 95% confidence estimates of the population mean mu **based on its sample only**
- Count the number of labs, where the population mean lies **outside their bounds**
- You should get ~50 labs out of 1000 labs

How I did it

- n=20; k_labs=1000;
- rand_table=2.*randn(n,k_labs)+3;
- sample_mean=mean(rand_table,1);
- CI_low=sample_mean-1.96.*2./sqrt(n);
- CI_high=sample_mean+1.96.*2./sqrt(n);
- k_above=sum(3>CI_high)
- k_below=sum(3<CI_low)
- figure; ndisp=100; errorbar(1:ndisp, sample_mean(1:ndisp), ones(ndisp,1).*1.96.*2./sqrt(n),'ko');
- hold on; plot(1:ndisp, 3.*ones(ndisp,1),'r-');

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

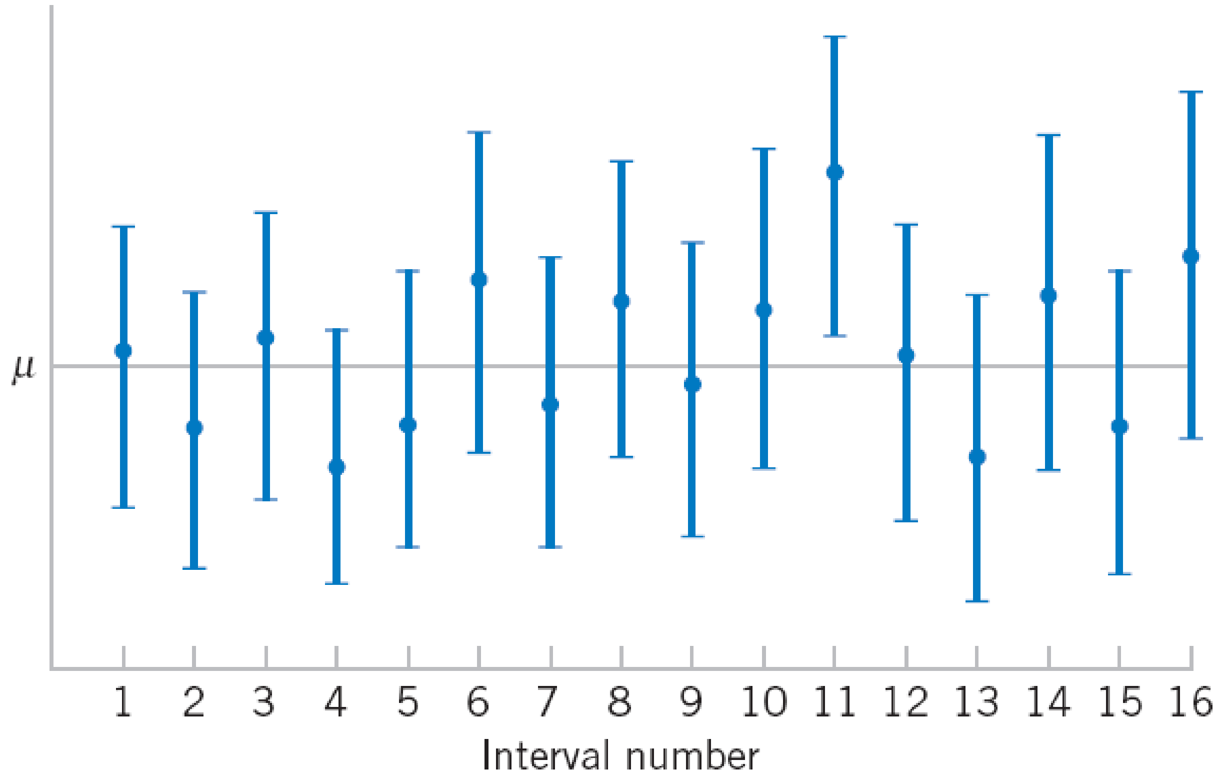


Figure 8-1 Repeated construction of a confidence interval for μ.

Figure 8-1 Repeated construction of a confidence interval for μ.

So far in estimating
confidence intervals for population mean μ
we assumed that the population variance σ^2
is known

Then (or when $n>>1$, say 20 and above)
one can use the Normal Distribution
to calculate confidence intervals

Q: What to do if the sample is small
& the population variance is **not known**?

A: Use the sample variance

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

but carefully:

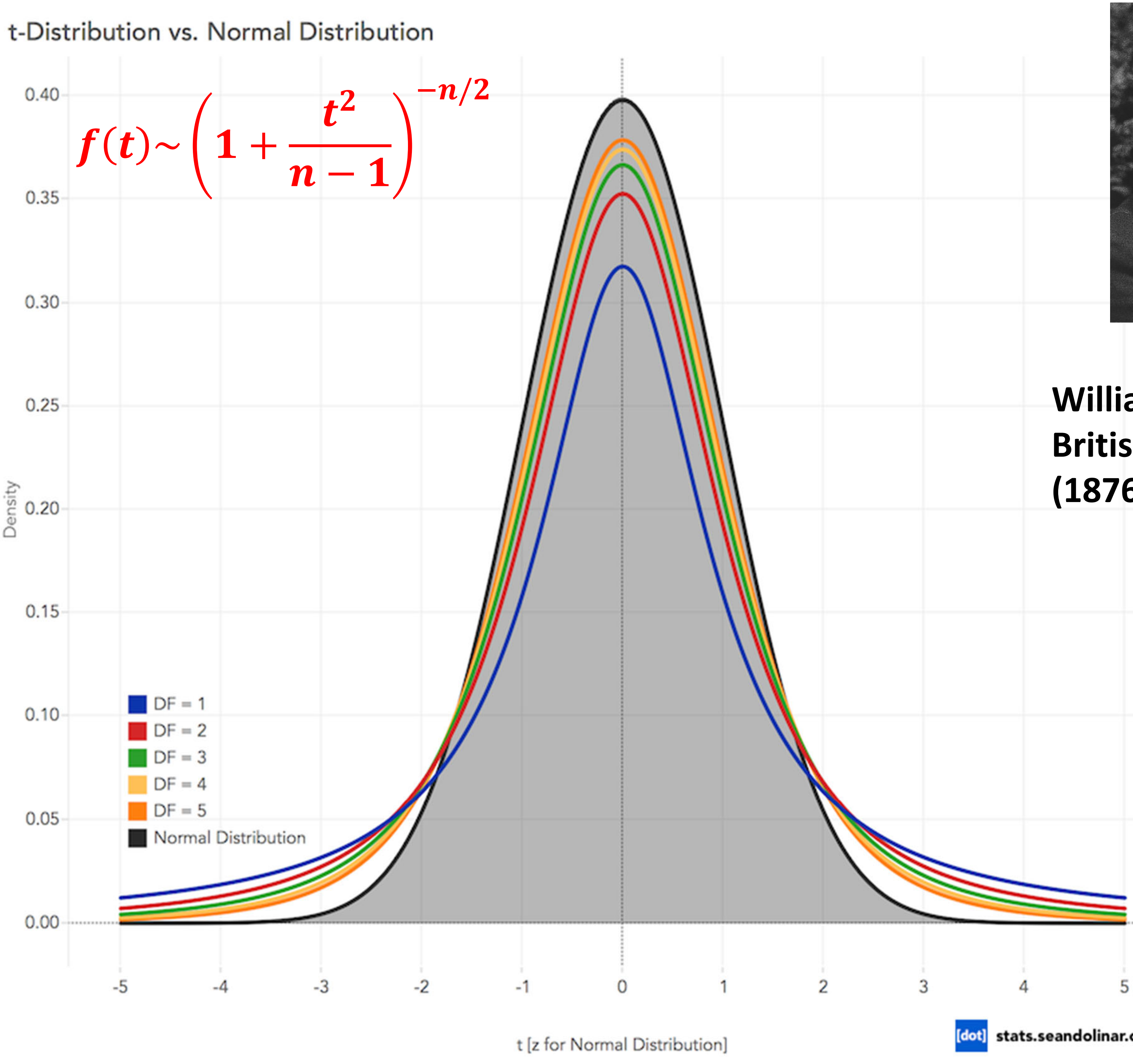
- Variable X has to be **normally distributed**
- **Student t-distribution** has to be used instead of the normal distribution (z-distribution).

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. However, after pleading with the brewery and explaining that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, he was allowed to publish them, but under a pseudonym ("Student"), to avoid difficulties with the rest of the staff. Thus, his most noteworthy achievement is now called Student's, rather than Gosset's, t-distribution.

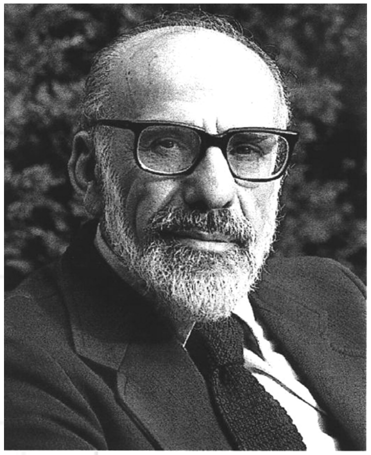
William Sealy Gosset
(13 June 1876 – 16 October 1937)
was an English statistician, chemist and
brewer who as Head Brewer of Guinness

Gosset had almost all his papers including “The probable error of a mean” (1908) published in Pearson's journal Biometrika under the pseudonym Student

Student's t-distribution



$$f(t) \sim \left(1+\frac{t^2}{n-1}\right)^{-n/2}$$



William Sealy Gosset
British statistician
(1876-1937)

Play with Mathematica notebook

http://demonstrations.wolfram.com/ComparingNormalAndStudentsTDistributions/

By Gary McClelland

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

$$\boldsymbol{\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}}$$

Student's *t* distribution

$$f(t) \sim \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

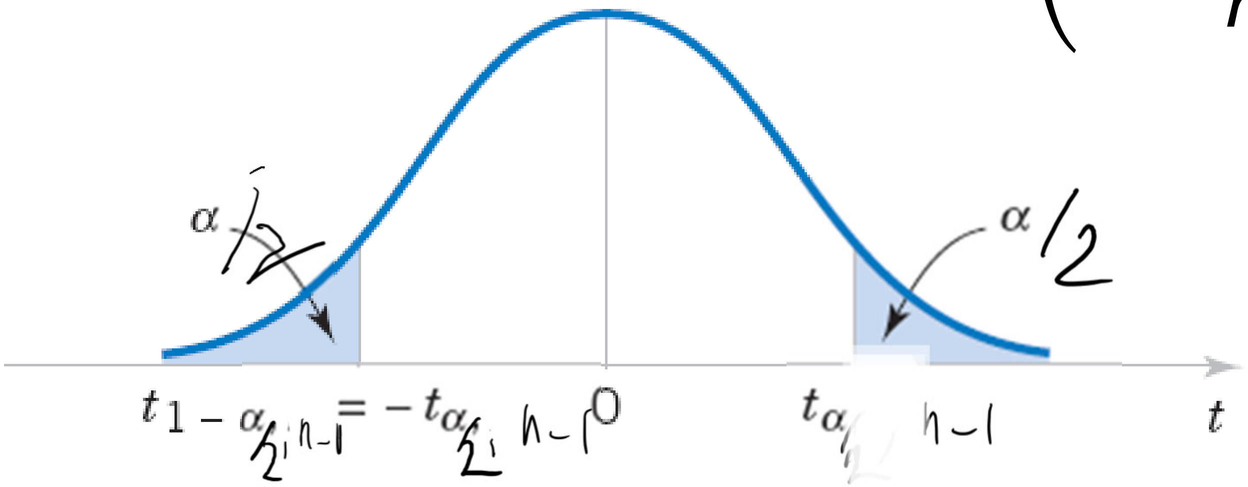
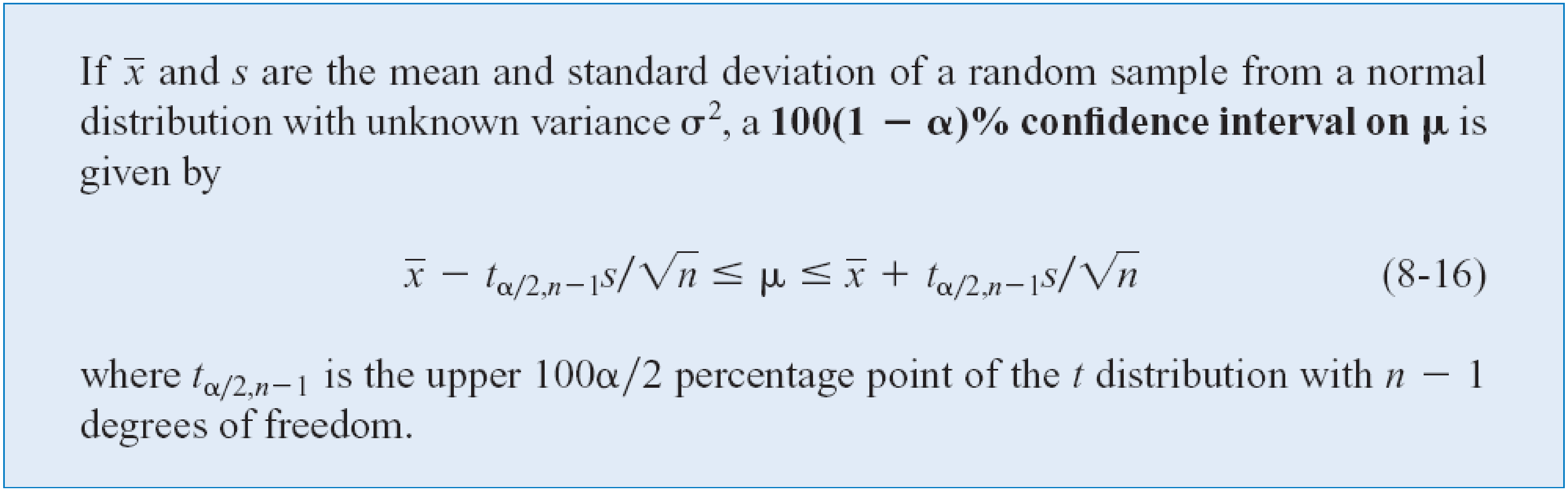


Figure 8-5 Percentage points of the *t* distribution.

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.2 The *t* Confidence Interval on μ (Eq. 8-16)



If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a **100(1 − α)% confidence interval on μ** is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \qquad (8\text{-}16)$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

One-sided confidence bounds on the mean are found by replacing $t_{\alpha/2,n-1}$ in Equation 8-16 with $t_{\alpha,n-1}$.

Confidence intervals for the population variance σ^2 based on the sample variance s^2

Confidence interval for the population variance σ^2

- Up until now we were calculating the confidence interval on the **population average μ**
- What if one wants to put **confidence interval on the population variance $\boldsymbol{\sigma^2}$**?
- We know an unbiased estimator of σ^2:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- How to determine the confidence interval?

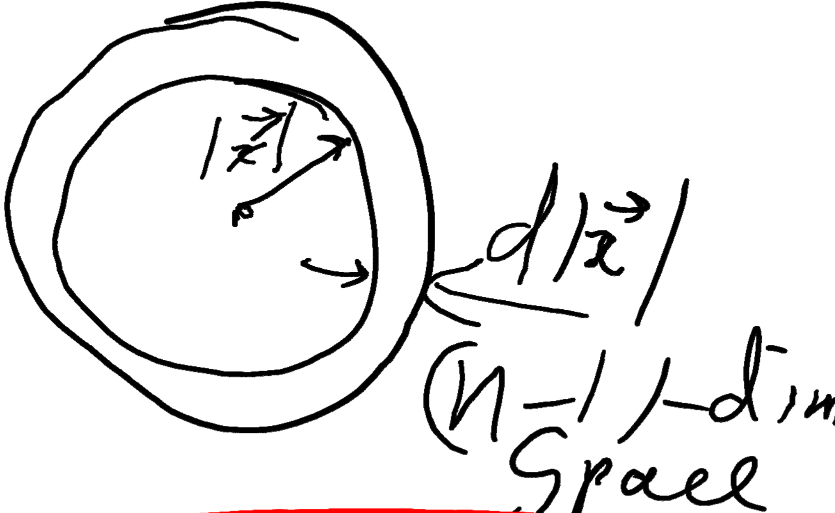
$\vec{X} = (x_1, x_2 \ldots x_n)$ $\qquad x_i \to x_i - \bar{x}$

$y = |\vec{x}|^2 = \sum x_i^2 = (n-1)s^2$ $\qquad \sum_{i=1}^{n} x_i = 0$

$$P(|\vec{x}|)\, d|\vec{x}| \sim \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) dx_i$$

(left the last one since $x_n = -\sum_{i=1}^{n-1} x_i$)

sphere area ~ $|\vec{x}|^{n-2}$



$|\vec{x}| = \sqrt{y}$

$d|\vec{x}| = \frac{1}{\sqrt{y}} dy$

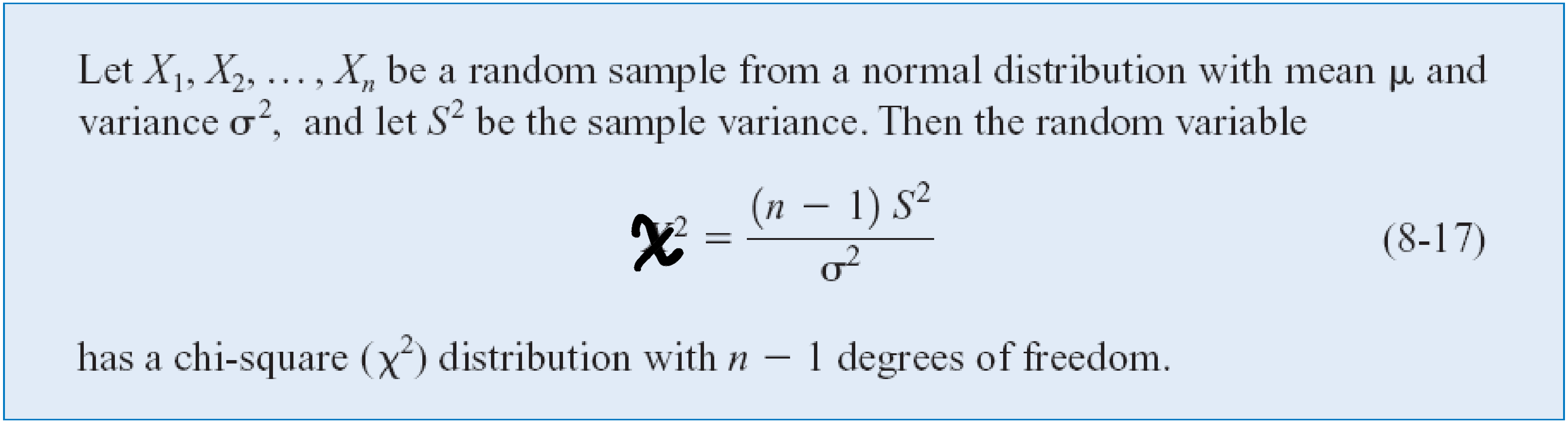
$\prod dx_i \sim |\vec{x}|^{n-2} d|\vec{x}|$

$$\boxed{P(y)\, dy = y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right) dy}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-17)



Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2} \qquad (8\text{-}17)$$

has a chi-square (χ^2) distribution with $n - 1$ degrees of freedom.

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$X=(n-1)S^2/\sigma^2$

We know n, S^2

want to estimate σ^2

$f(x,n) \sim x^{(n-1)/2-1}\exp(-x/2)$

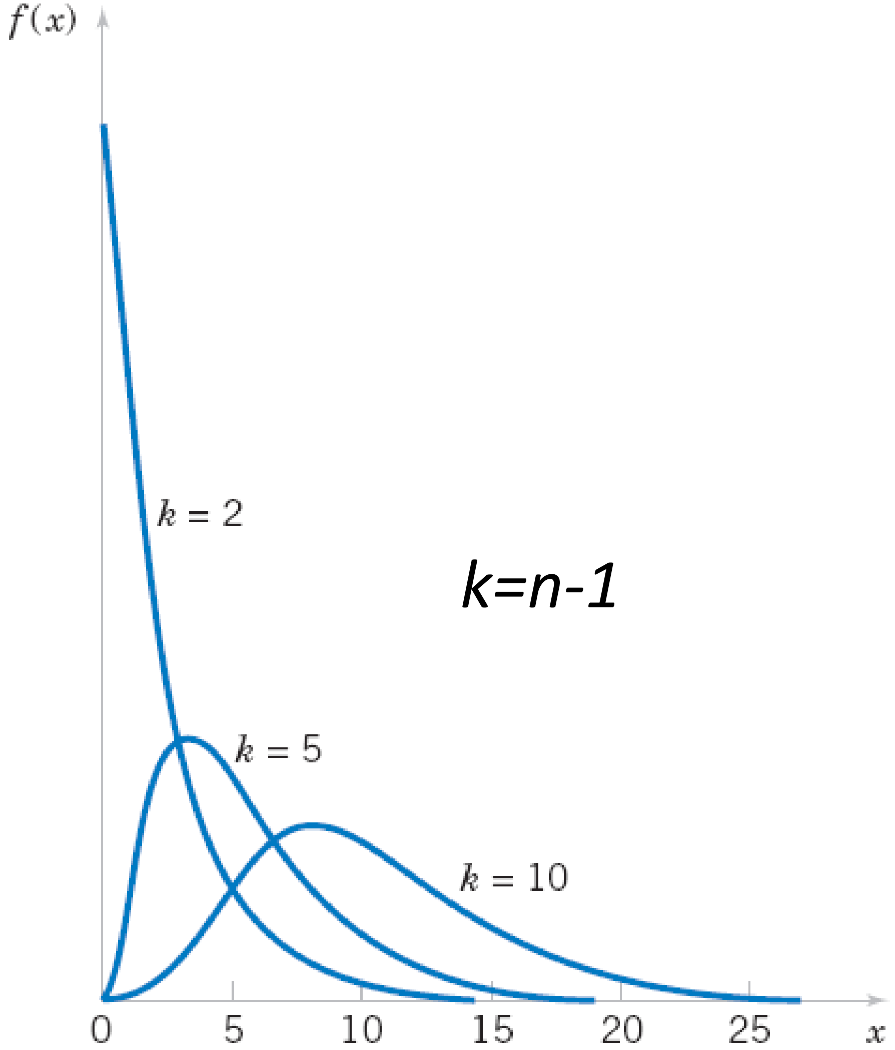
It is just Gamma PDF with *$r=(n-1)/2$,* and $\lambda=1/2$

Mean value:

$n-1$

Standard deviation:

$$\sqrt{2(n-1)}$$

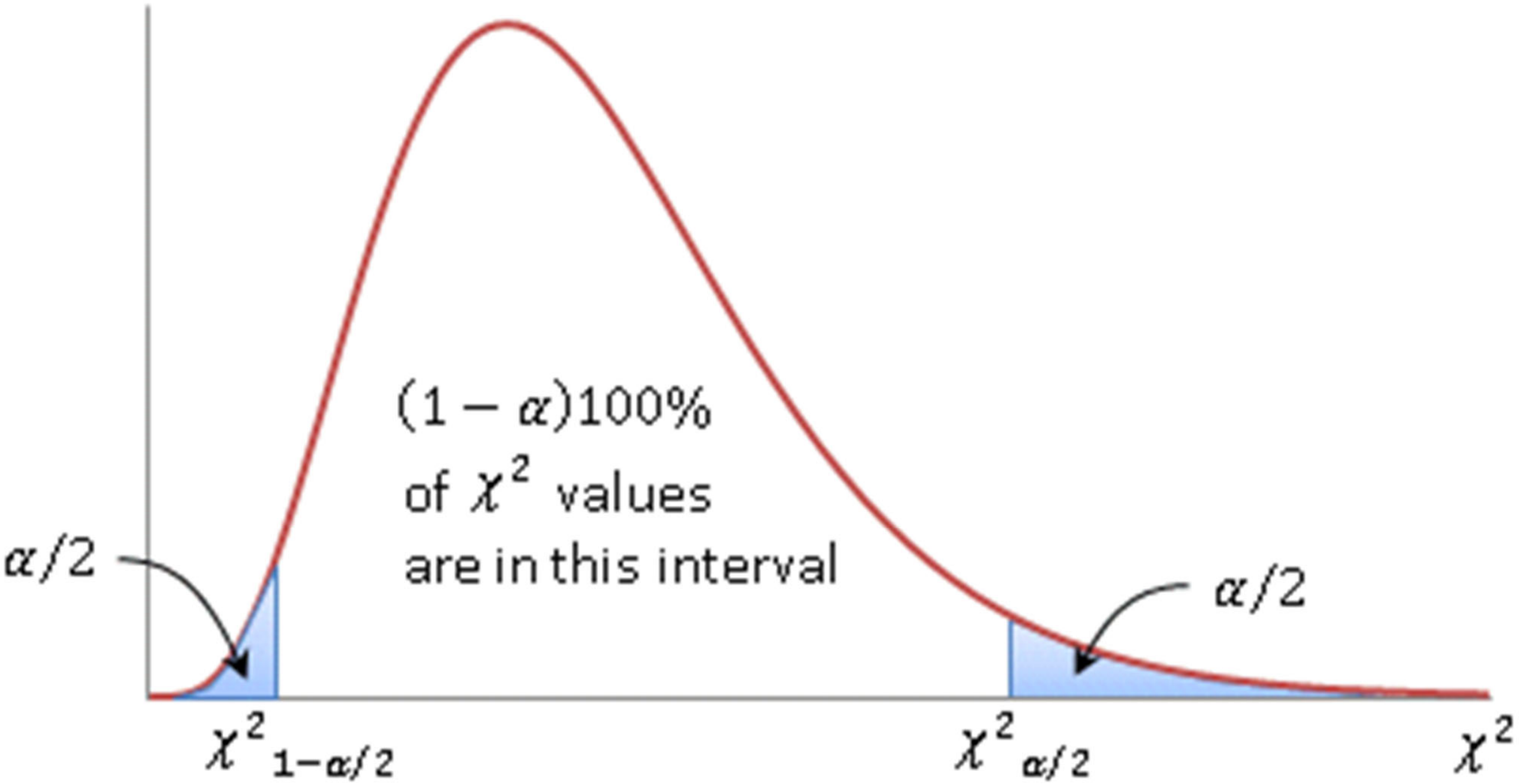


k=n-1

Play with Mathematica notebook

http://demonstrations.wolfram.com/ChiSquaredDistributionAndTheCentralLimitTheorem/

By Peter Falloon



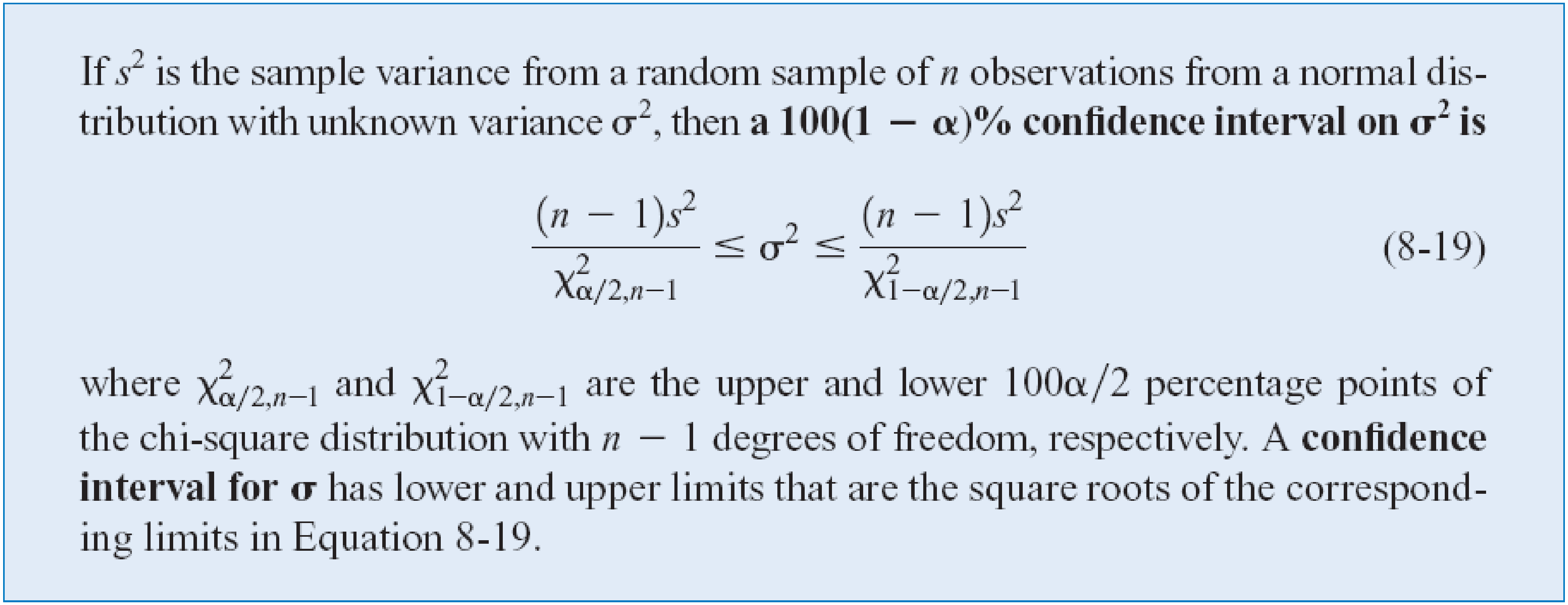
$$\chi^2{}_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2{}_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2{}_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2{}_{1-\alpha/2}}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-19)

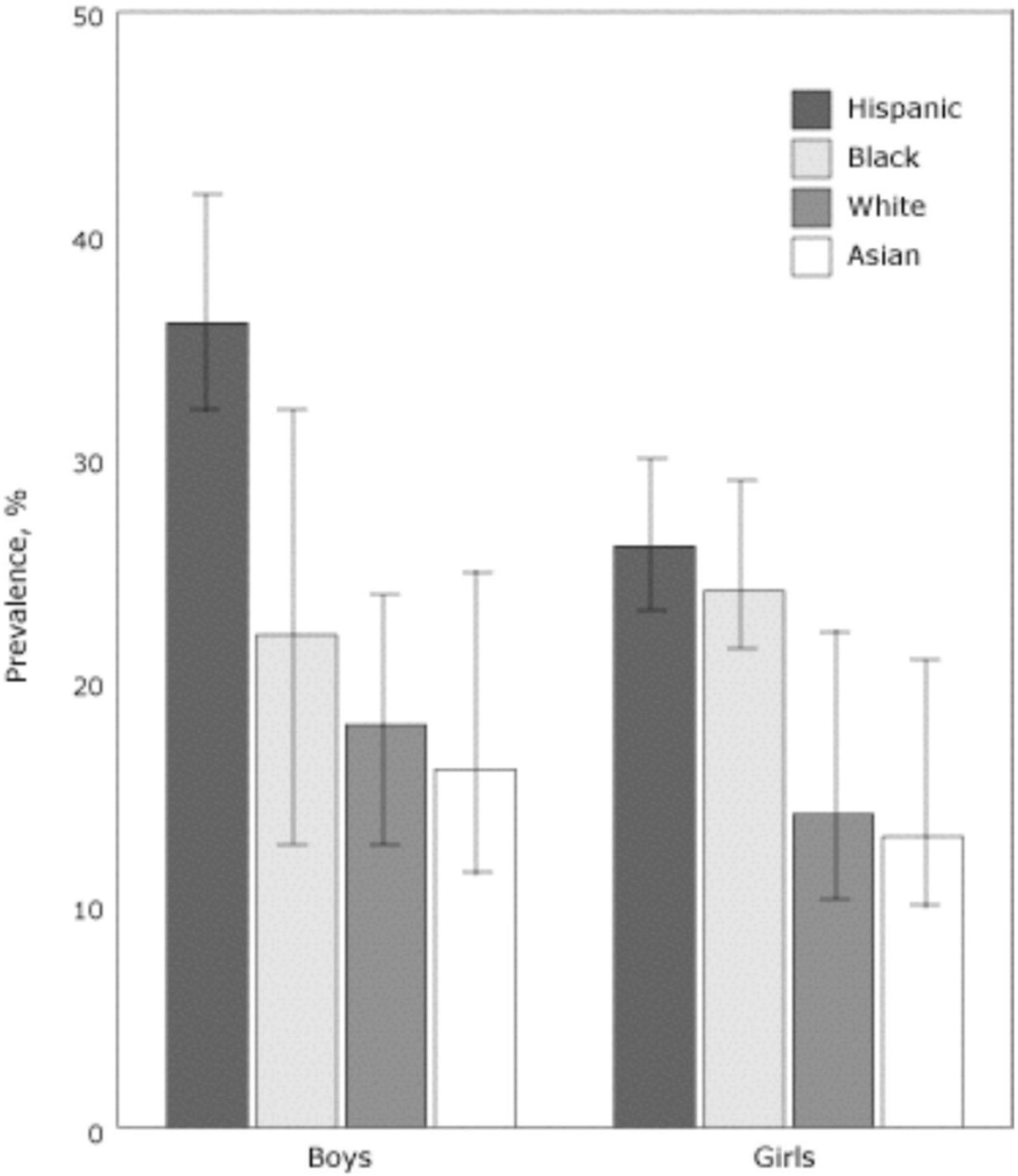


If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a 100(1 − α)% confidence interval on σ^2 is**

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{8-19}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. A **confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.

Confidence estimates of the population proportion



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

Collect a sample of BMI values

Obese means BMI>30

What do those bars actually mean?

Large sample confidence estimate of population proportion

- Want to know the fraction *p* of the population that belongs to a class, e.g., the class “obese” kids defined by BMI>30.
- Each variable is a Bernoulli trial with one parameter p.
  We can use moments or MLE estimator to estimate p
- Both give the same estimate: sample fraction *$\hat{p}$=(# of obese kids in the sample)/(sample size n)*
- How to put confidence bounds on *p* based on $\hat{p}$
- # of obese kids in the sample follows the binomial distribution:
  “success” = sampled kid is obese : -(
  p – probability of success, *1-p* – failure
- Expected # of successes is *np* → Expected fraction of successes is *p*
- Standard deviation of # of successes is $\sqrt{np(1-p)}$→

Standard deviation of fraction of successes is $\sqrt{p(1-p)/n}$

8-5 A Large-Sample Confidence Interval For a Population Proportion

Normal Approximation for Binomial Proportion

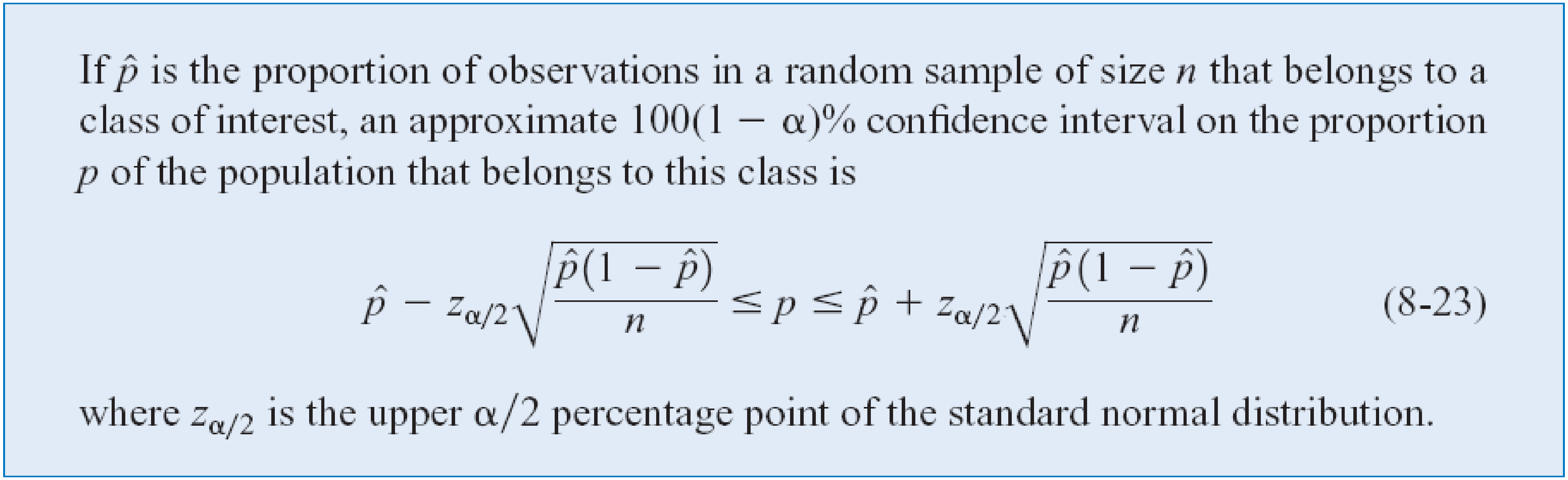
If n is large, the distribution of

$$Z = \frac{\overset{\simeq n\hat{p}}{X} - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \simeq \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately standard normal.

The quantity $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard error of the point estimator . $\hat{P}$

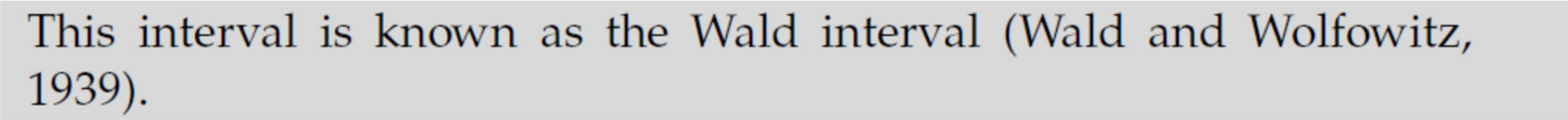
8-5 A Large-Sample Confidence Interval For a Population Proportion (Eq. 8-23)



If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate $100(1-\alpha)\%$ confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{8-23}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.



This interval is known as the Wald interval (Wald and Wolfowitz, 1939).

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

“To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to estimate these probabilities from a finite sample and how to set confidence interval on these estimates?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

For blue M&Ms $p = 0.24$

$$1.96\sqrt{\frac{0.24(1-0.24)}{n}} < 0.04$$

$n > \left(\frac{1.96}{0.04}\right)^2 0.24\times(1-0.24) = 438$ M&Ms or

~2 × 7oz bags with 210 candies each

For red M&Ms $p = 0.13$

$n > \left(\frac{1.96}{0.04}\right)^2 \times 0.13 \times (1-0\times13) \simeq 271$ M&Ms or

~1 × 7oz bag