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**?

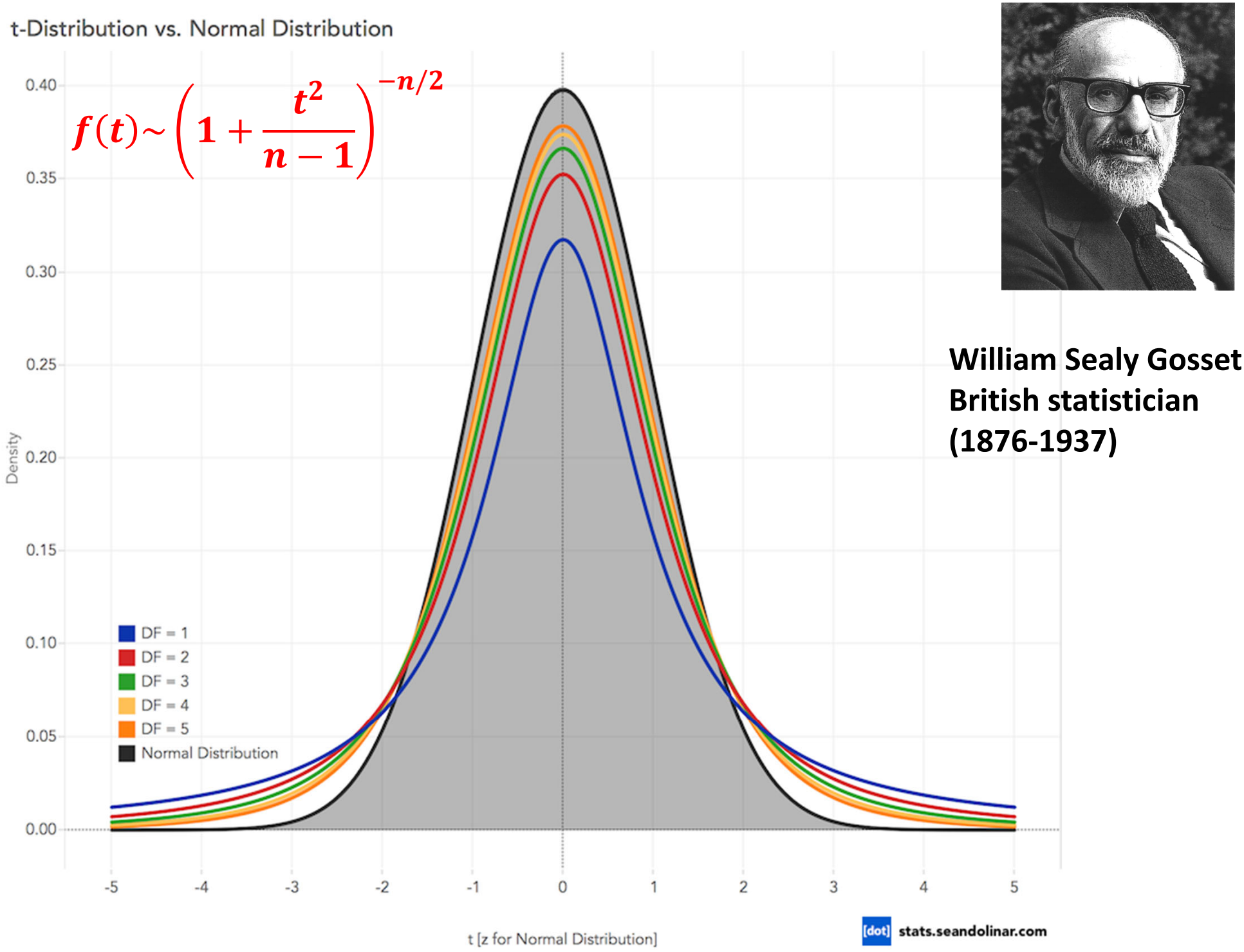
A: Use the sample variance

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

but carefully:

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

Student's t-distribution



William Sealy Gosset
British statistician
(1876-1937)

Play with Mathematica notebook

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

By Gary McClelland

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

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

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

Student's *t* distribution

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

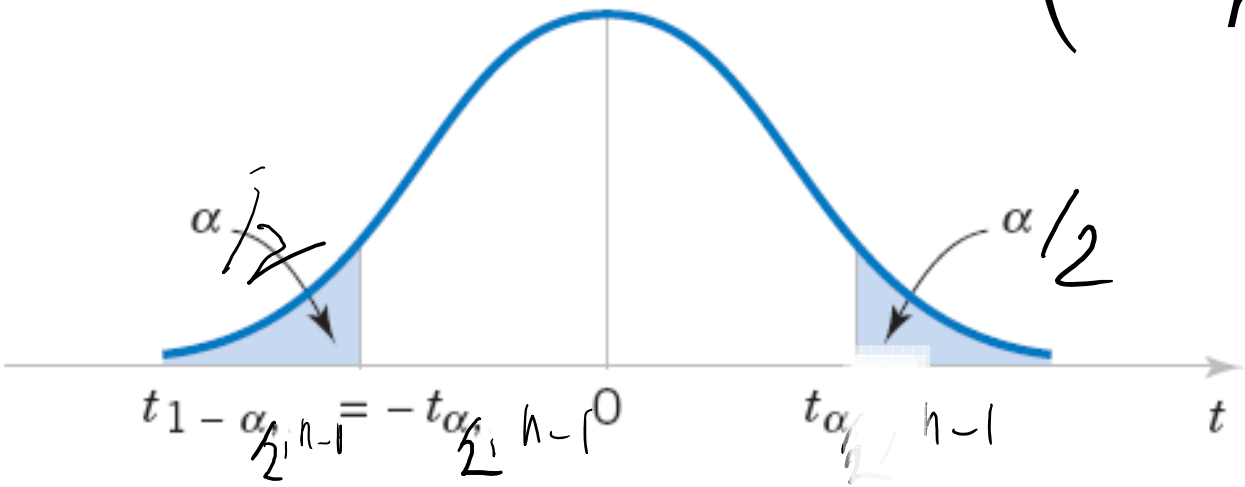


Figure 8-5 Percentage points of the *t* distribution.

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 σ^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?

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

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

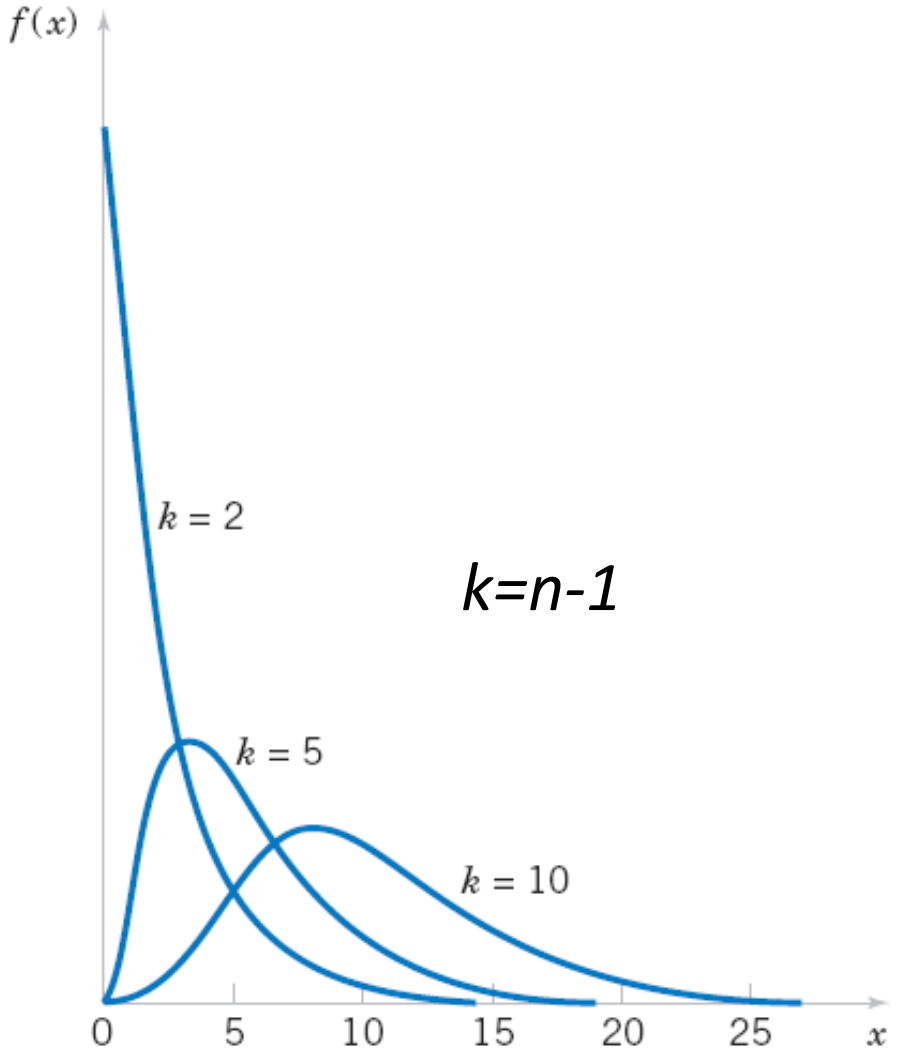
It is just Gamma PDF with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

$n-1$

Standard deviation:

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



$k=n-1$

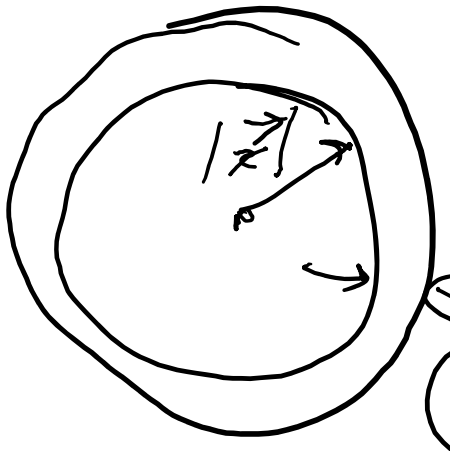
$\vec{X} = (x_1, x_2 \dots x_n)$ $\quad x_i \to x_i - \bar{x}$

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

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

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

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



$(n-1)$-dim Space

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

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

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

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

Number of degrees of freedom $k = n - 1$

Let $X = (X_1, \dots, X_n)$ iid $\mathcal{N}(0,1)$, $\bar{X} = \frac{1}{n}\sum X_i$, and consider the **residual sum of squares**

$$q = \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

Geometrically, decompose $\mathbb{R}^n$ into the span of $u = \frac{1}{\sqrt{n}}(1, \dots, 1)$ and its orthogonal complement $u^\perp$ (dimension $n - 1$). There exists an orthogonal matrix U with first column u, so

$$Y = U^\top X \sim \mathcal{N}_n(0, I_n), \quad Y_1 = \langle X, u \rangle = \sqrt{n}\,\bar{X}, \quad (Y_2, \dots, Y_n) \in u^\perp.$$

$$q = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{j=2}^{n} Y_j^2 = \| Y_\perp \|^2,$$

where $Y_\perp = (Y_2, \dots, Y_n) \sim \mathcal{N}_{n-1}(0, I_{n-1})$ and is **independent** of Y_1.

Now integrate in the $(n-1)$-dimensional space $u^\perp$. In $\mathbb{R}^{n-1}$, the surface area of the $(n-2)$-sphere of radius r is

$$S_{n-2}(r) = \frac{2\pi^{(n-1)/2}}{\Gamma(n - 1/2)}\, r^{n-2}.$$

The radial density for $R = \| Y_\perp \|$ is

$$f_R(r) = \frac{S_{n-2}(r)}{(2\pi)^{(n-1)/2}}\, e^{-r^2/2} = \frac{2^{1-(n-1)/2}}{\Gamma(n - 1/2)}\, r^{n-2} e^{-r^2/2}, \quad r > 0.$$

With $q = r^2$ and $dr = \frac{1}{2\sqrt{q}}\, dq$,

$$f_Q(q) = f_R(\sqrt{q})\, \frac{1}{2\sqrt{q}} = \frac{1}{2^{(n-1)/2}\Gamma(n - 1/2)}\, q^{\frac{n-1}{2}-1} e^{-q/2}, \quad q > 0.$$

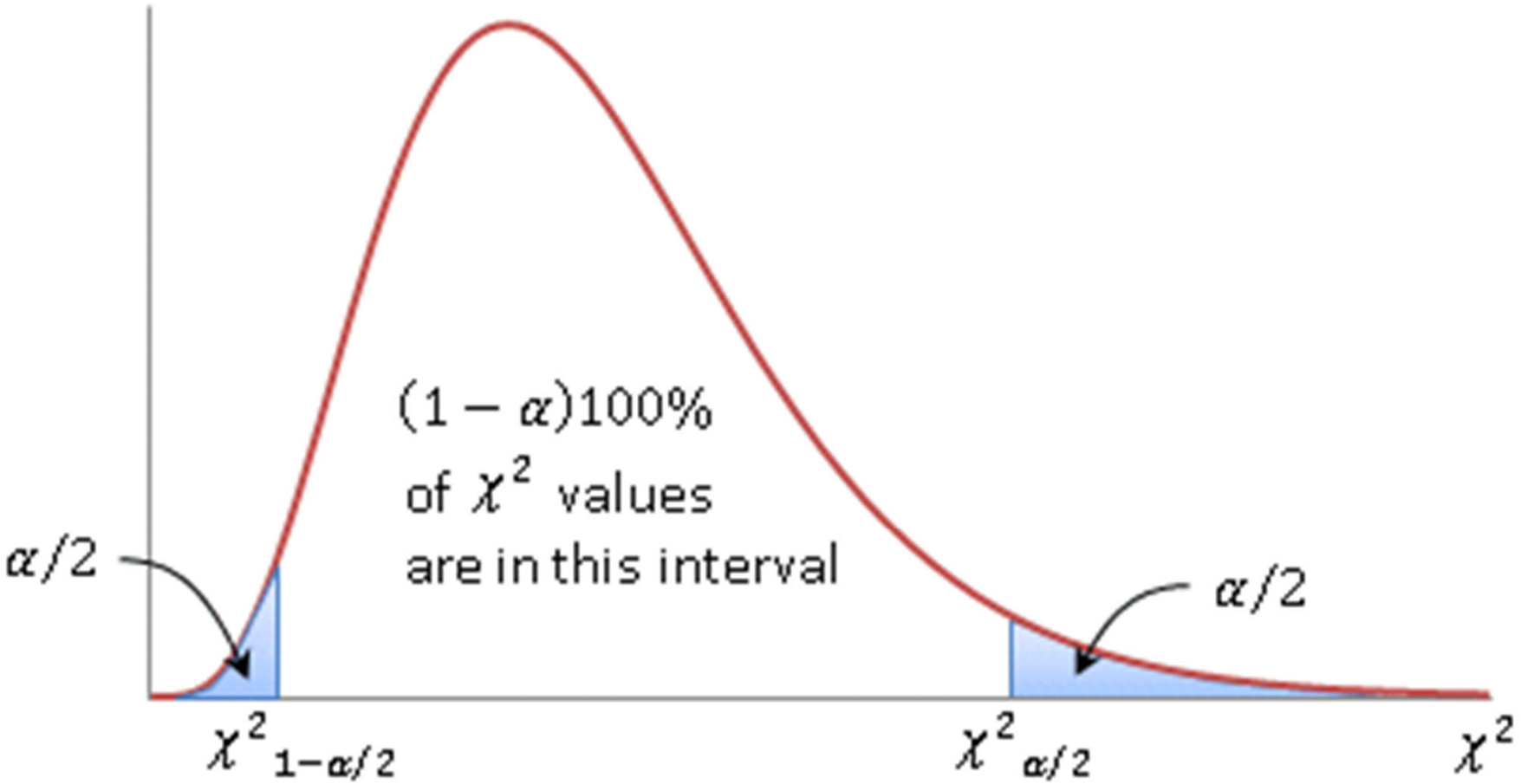
That is exactly χ^2_k with $k = n - 1$:

$$Q \sim \chi^2_{n-1}, \qquad f(q) = \frac{1}{2^{k/2}\Gamma(k/2)}\, q^{k/2-1} e^{-q/2}, \; k = n - 1.$$

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)

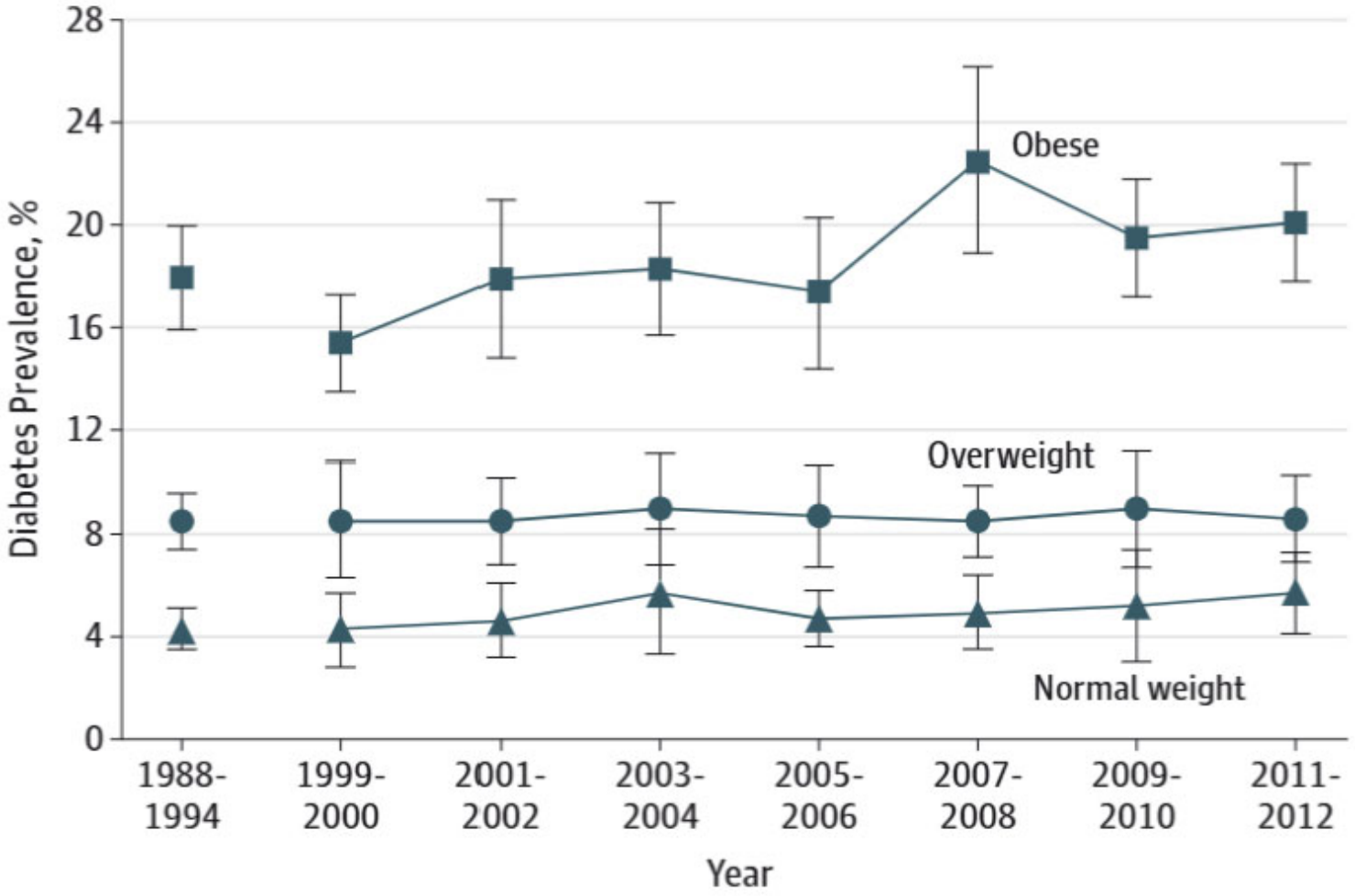
If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a $100(1-\alpha)\%$ confidence interval on σ^2 is**

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

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

Figure 2. US Trends in Diabetes Prevalence per 100 Adults Aged 20 Years or Older by BMI Category



No. of participants								
Obese	2324	727	732	815	820	1137	1302	1075
Overweight	2942	724	878	784	694	949	1009	852
Normal weight	3025	645	699	624	604	726	762	785

Menke A, Casagrande S, Geiss L, Cowie CC JAMA. 2015;314: 1021–1029

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 "people with diabetes"
- 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 people with diabetes in the sample)/(sample size n)*
- How to put confidence bounds on *p* based on $\hat{p}$
- Each participants in the sample is a Bernoulli trial: "success" = sampled participant has diabetes :-(
- Standard deviation of # of successes in n Bernoulli trials is given by Binomial distribution $\sqrt{n\,\hat{p}(1-\hat{p})}$
- Standard error of the fraction of successes is $\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

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

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

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 $p = 0.24$

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

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

$\sim$ 2 x 7oz bags with 210 candies each

For red M&Ms $p = 0.13$

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

$\sim$ 1 x 7oz bag