CSCI 3022: Intro to Data Science

Lecture 18: Statistical Inference with Small Samples

Rachel Cox

Department of Computer Science

William Sealy Gosset

Announcements & Reminders

HW3 - Due Monday March 30th at 11:59pm

HW4 - Due Friday Apr. 3 at 11:59pm

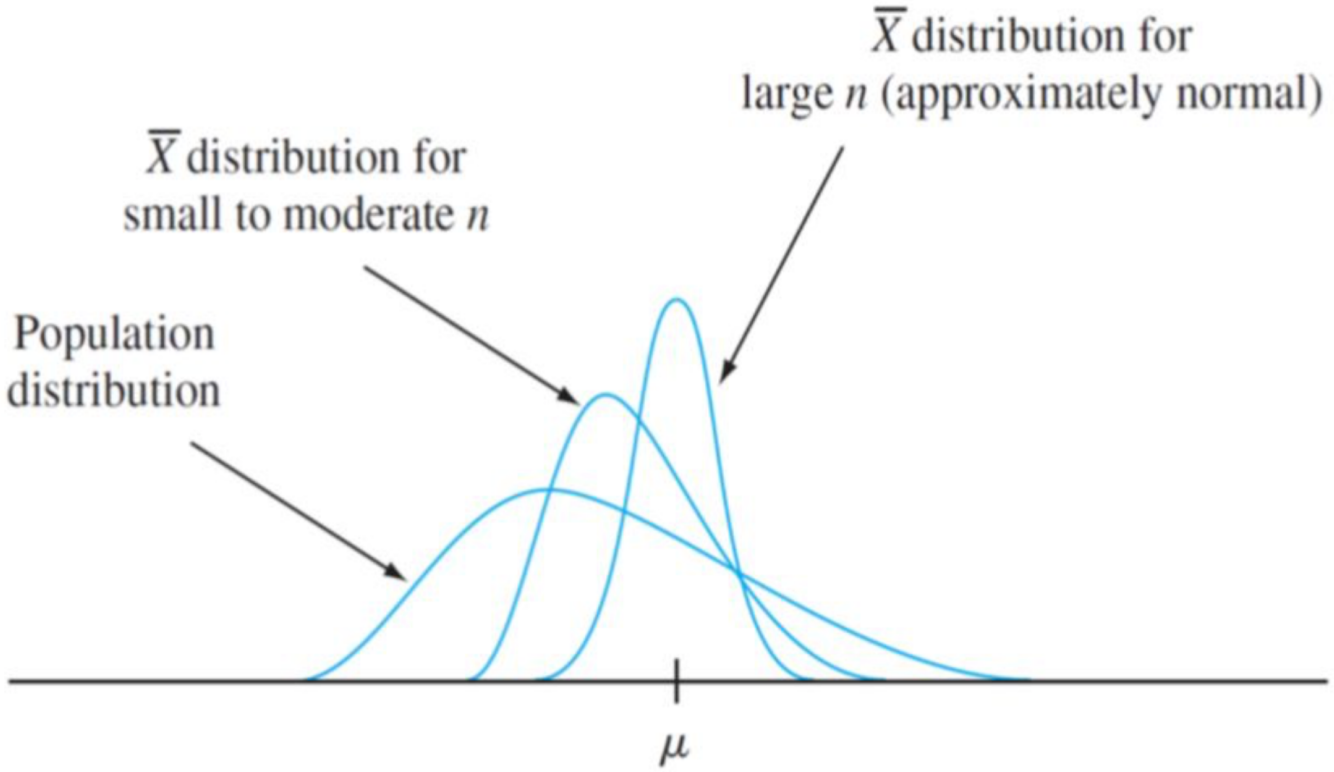
HW5 - last HW! - will be posted at the end of the week

What will we learn today?

- ❑ Small samples – what distributions to use (instead of Standard Normal) when our sample is too small for the CLT to kick in
- ❑ T-distributions – small sample CI/hypothesis testing for the mean
- ❑ Chi-Squared distributions – small sample CI/hypothesis testing for variance

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 27*

Distribution of the Sample Mean



Visualization from onlinestatbook.com

The Central Limit Theorem

The Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be iid draws from some distribution. Then, as n becomes large

$n \geq 30$

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

From the book: Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each X_i. For $n \geq 1$, let Z_n be defined by

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}$$

Then for any number a, $\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a)$, where Φ is the distribution function of the N(0, 1) distribution.

- The distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

Variance

Sample Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$

default ddof = 0 $\div n$

ddof = 1 $\div n-1$ *

- Some intuition: Suppose you want to test the population mean with a sample of 10. It must be that $\sum_{i=1}^{10} x_i = n \cdot \bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n}$$

- The first 9 data values can be anything – *freely* chosen. The 10th data value is determined by the first 9 data values.

- You have 10-1 = 9 "degrees of freedom"

* Khan Academy video with neat visualization of s^2 computed with denominators of $n, n-1,$ and $n-2$.

Bessel's Correction

We want an "unbiased estimator"

$E[s^2] = E\left[\frac{\sum (x_i - \bar{x})^2}{n-1}\right]$ what if we divide by n instead of $n-1$

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

$= \frac{1}{n-1} E[\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)]$

$= \frac{1}{n-1} E[\sum_i x_i^2 - 2\bar{x}\underbrace{\sum_i x_i}_{n\bar{x}} + \sum_i \bar{x}^2]$

$= \frac{1}{n-1} E[\sum_i x_i^2 - 2\bar{x}^2 \cdot n + \sum_i \bar{x}^2]$

$= \frac{1}{n-1}\left(E[\sum_i x_i^2] - 2n\,E[\bar{x}^2] + E[\sum_i \bar{x}^2]\right)$

$= \frac{1}{n-1}\left(\sum_i E[x_i^2] - 2n\,E[\bar{x}^2] + n\,E[\bar{x}^2]\right)$

$= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - 2n\left(\frac{\sigma^2}{n}+\mu^2\right) + n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)$

$= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)$

$= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) = \frac{1}{n-1}(\sigma^2(n-1)) = \frac{n-1}{n-1}\sigma^2 = \sigma^2$

- $E[s^2] = \sigma^2$
- $E[\bar{x}] = \mu$

Recall:

- $\mathrm{Var}(x) = \sigma^2 = E[x^2] - (E[x])^2$
  $= E[x^2] - \mu^2$
  $\Rightarrow E[x^2] = \sigma^2 + \mu^2$
- CLT: $\bar{x} \sim N(\mu, \sigma^2/n)$
  $E[\bar{x}^2] = \frac{\sigma^2}{n} + \mu^2$

Statistical Inference – Distribution Type and Sample Size

Statistical inference for population mean when data are normal and n is large and …

σ is known: $\bar{X} \overset{CLT}{\sim} N(\mu, \sigma^2/n)$

σ is unknown: $s^2 \simeq \sigma^2 \rightarrow \bar{X} \sim N(\mu, s^2/n)$

Statistical inference for population mean when data are NOT normal and n is large and …

σ is known: $\bar{X} \sim N(\mu, \sigma^2/n)$ CLT.

σ is unknown: $\bar{X} \sim N(\mu, s^2/n)$

Statistical Inference – Distribution Type and Sample Size

Statistical inference for population mean when data are normal and n is small and …

σ is known: $\bar{X} \sim N(\mu, \sigma^2/n)$

σ is unknown: $\bar{X}$ is not necessarily normally distributed → use the student t-distribution ★ Today's focus

Statistical inference for population mean when data are NOT normal and n is small and …

σ is known:

σ is unknown:

} Bootstrap – next lecture

Statistical Inference – Distribution Type and Sample Size

To summarize:

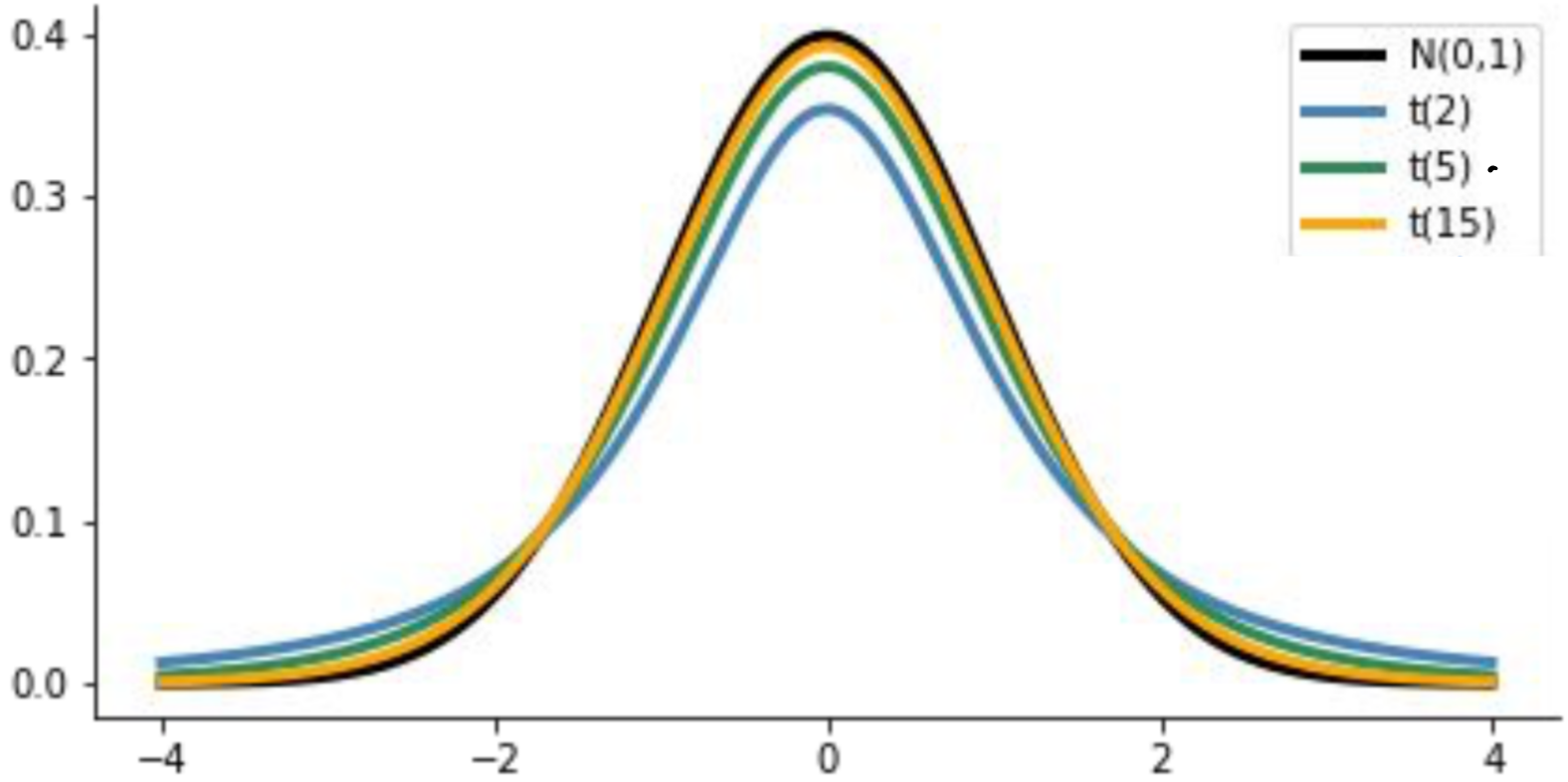
	$n \geq 30$	$n < 30$
Normal data, known σ	Z-test	Z-test
Normal data, unknown σ	Z-test	t-test
Non-normal data, known σ	Z-test	Bootstrap
Non-normal data, unknown σ	Z-test	Bootstrap

Properties of t–distributions

When $\bar{X}$ is the sample mean of a random sample of size n from a normal distribution with mean μ, the random variable t follows a probability distribution called a **t-distribution** with parameter $\nu = n - 1$ degrees of freedom.

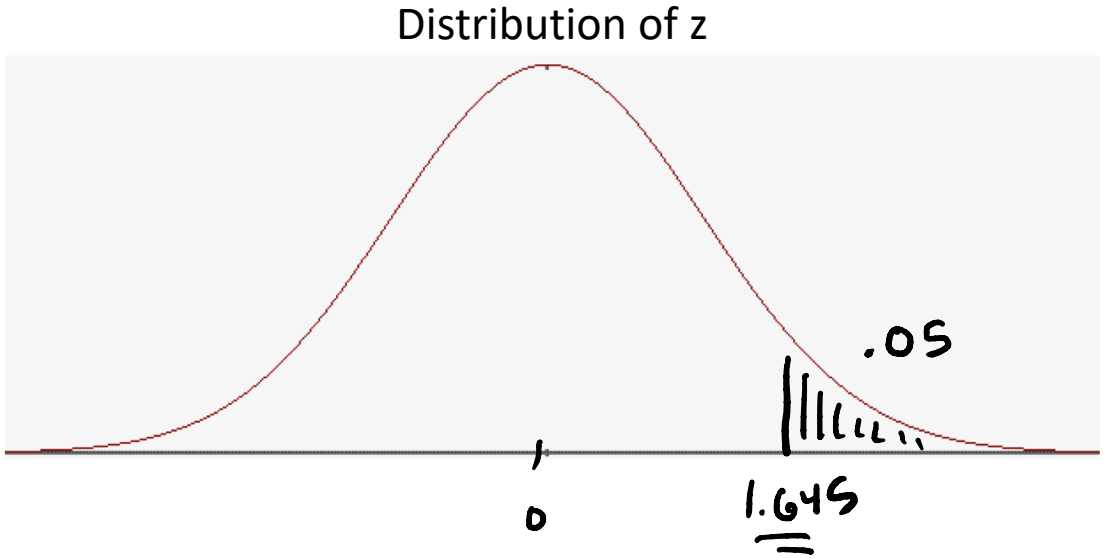
$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

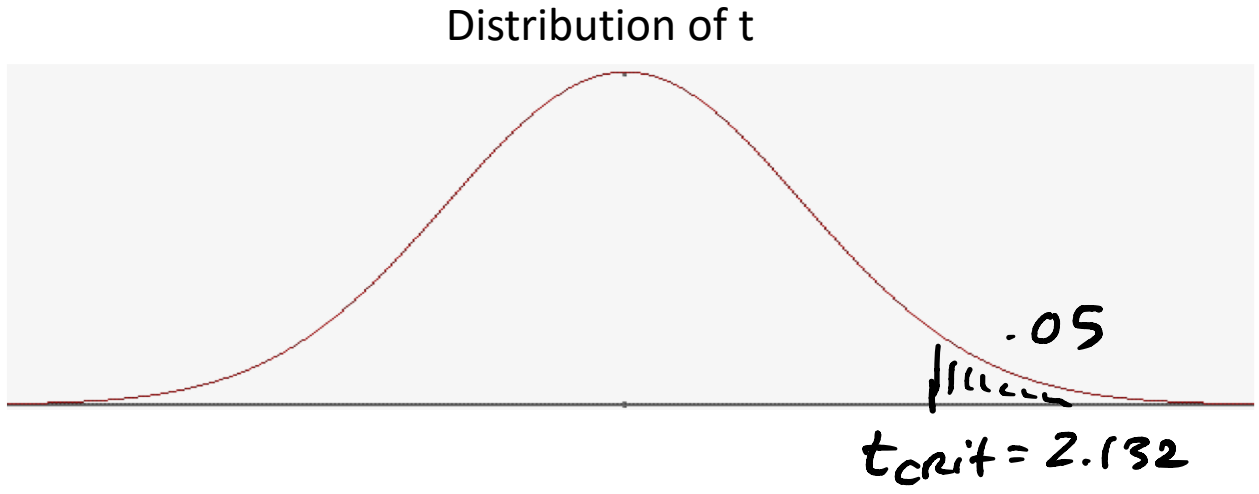
- Each t_ν curve is bell-shaped, centered at 0.
- Each t_ν curve is more spread out than the standard normal distribution.
- As ν increases, the spread of the corresponding t_ν curve decreases.
- As $\nu \to \infty$ the sequence of t_ν curves approaches the standard normal curve.



The t–critical value

- We can extend all of our inferential mechanics to the small-sample case by introducing the so-called t-critical value, which we denote as $t_{\alpha,\nu}$ → stats.t.ppf(1−α, df=ν)
- The t-critical value $t_{\alpha,\nu}$, is the point such that the area under the t_ν- curve to the right of $t_{\alpha,\nu}$ is equal to α.





- Because student's t distribution is more spread out, the critical values of t for a given level of significance are larger in magnitude than the corresponding z critical values

The t–critical value

Example: $t_{0.05,6}$ is the t-critical value that captures the upper-tail area of 0.05 (5%) under the t-curve with 6 degrees of freedom.

What is the sample size?

$\nu = n - 1$

$\nu = 6 \quad \Rightarrow \quad n = \nu + 1 = 7$

$n = 7$

The t–confidence interval for the mean

Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from a random sample of size n, from a normal population with mean μ.

The $100 \cdot (1-\alpha)\%$ t-confidence interval for the mean μ is given by:

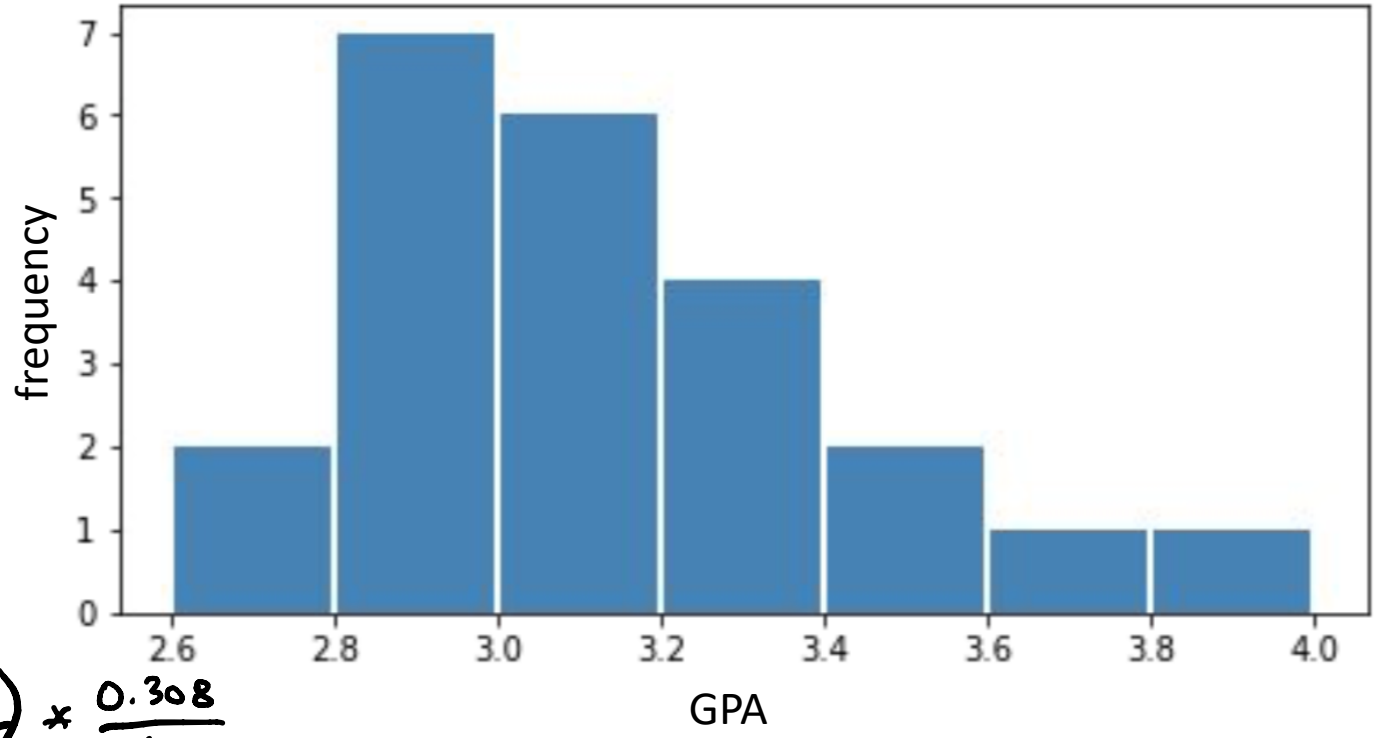
z CI: $\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$

t CI: $\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$

Critical values for t: stats.t.ppf$\left(1 - \frac{\alpha}{2}, df = n - 1\right)$

The t–confidence interval for the mean

Example: Suppose the GPAs for 23 students have the histogram shown here. The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Find a 90% CI for the mean GPA.



$n = 23$

$\nu = 23 - 1 = 22$

$\bar{x} = 3.146$

$s = 0.308$

$\alpha = 0.90$

use the student t-distribution.

$$\bar{x} \pm t_{0.05,22} * \frac{s}{\sqrt{n}}$$

$$= 3.146 \pm \text{stats.t.ppf}(.95, df = 22) * \frac{0.308}{\sqrt{23}}$$

$$= 3.146 \pm 1.717 * \frac{0.308}{\sqrt{23}}$$

CI: $[3.0357, 3.2563]$

The t–test, critical regions, and p–values

Alternative hypothesis	Critical region level α test	p-value level α test
$H_1: \theta > \theta_0$	$t \geq t_{\alpha,\nu}$	$P(T \geq t \mid H_0 = \text{true}) \leq \alpha$
$H_1: \theta < \theta_0$	$t \leq t_{\alpha,\nu}$	$P(T \leq t \mid H_0 = \text{true}) \leq \alpha$
$H_1: \theta \neq \theta_0$	$(t \geq t_{\alpha/2,\nu})$ or $(t \leq -t_{\alpha/2,\nu})$	$2 \cdot P(T \leq -\lvert t \rvert \mid H_0 = \text{true}) \leq \alpha$

upper tailed test

lower-tailed test

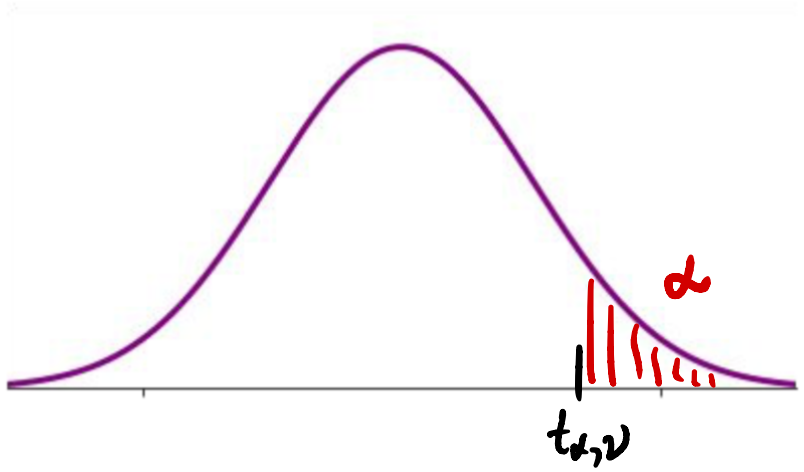
two tailed test

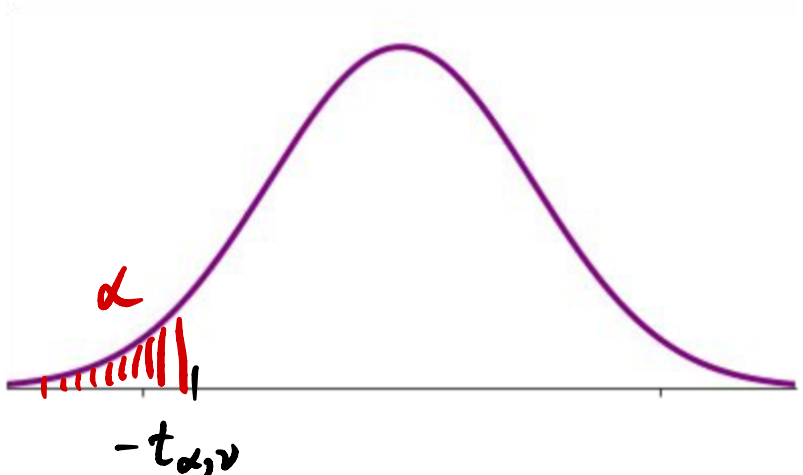
pvalue

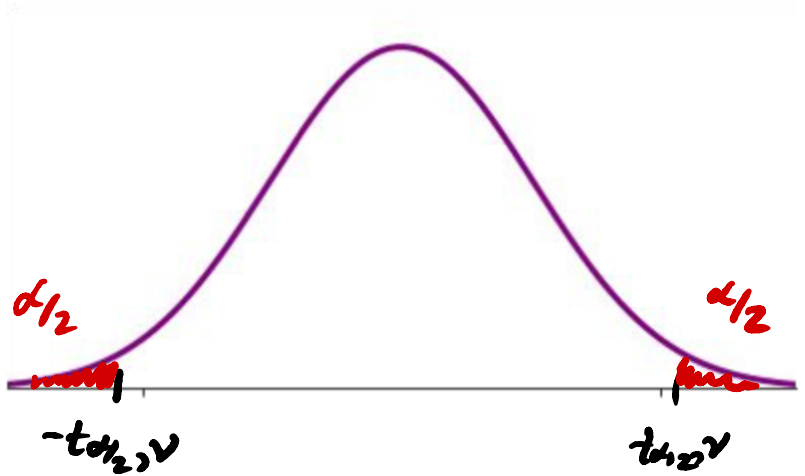
$p\text{-value} = 1 - \text{cdf}(t_{stat}, \nu)$

$p\text{-value} = \text{cdf}(t_{stat}, \nu)$

$p\text{-value} = 2 * \text{cdf}(-|t_{stat}|, \nu)$

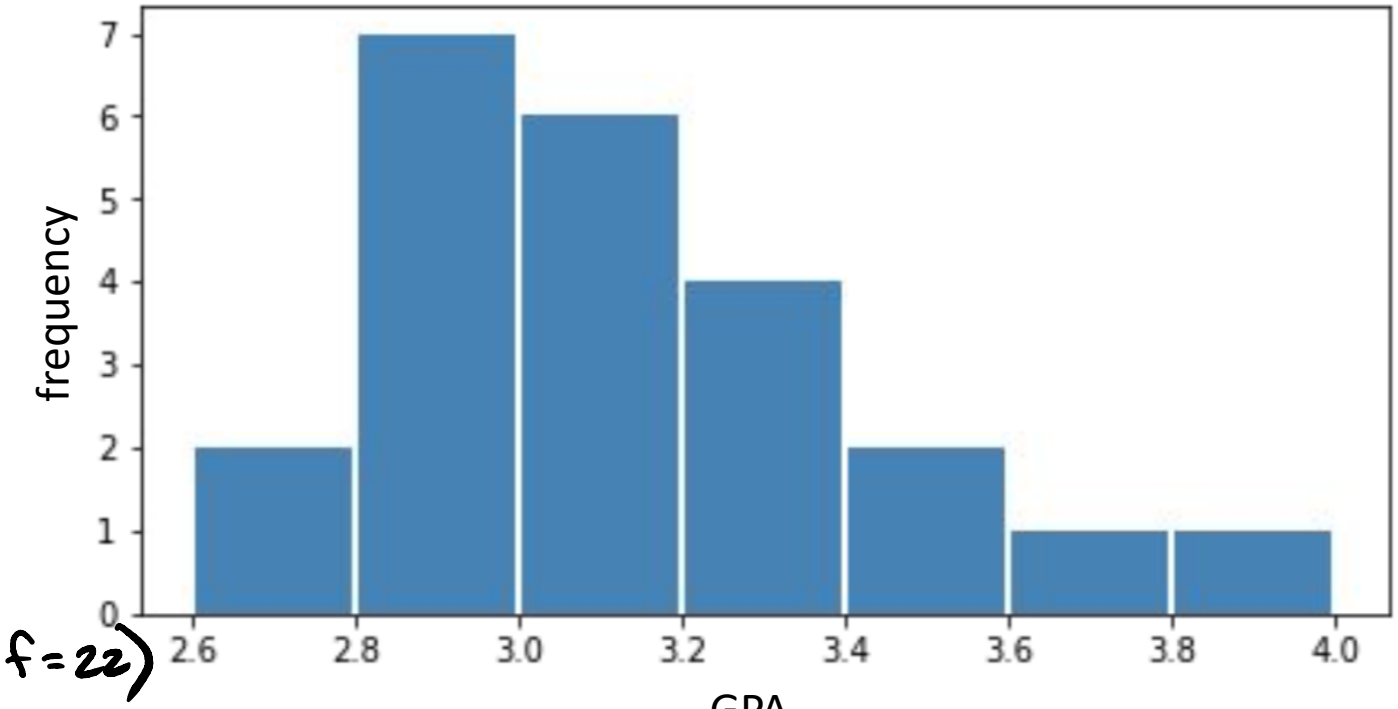






t–test for the mean, using p–values

Example: Suppose the GPAs for 23 students have the histogram shown here. The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Determine if there is sufficient evidence to conclude at the 0.10 (10%) significance level that the mean GPA is not equal to 3.30.



$H_0: \mu = 3.30$

$H_1: \mu \neq 3.30$

$$t_{stat} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3.146 - 3.30}{0.308/\sqrt{23}}$$

$$\cong -2.3979$$

p-value = 2 * stats.t.cdf(−2.3979, df=22)

$= 2 * 0.0127$

$= 0.0254$

$< 0.1 = \alpha$

So we've found p-value $< \alpha$

$\Rightarrow$ Reject the null hypothesis

$\Rightarrow$ There is sufficient evidence at the 10% sig level that $\mu \neq 3.30$

Inference for Variances

We've talked about confidence intervals for the mean and for proportions.

What does the sampling distribution of the variance look like when population is normally distributed?

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Define the sample variance as:

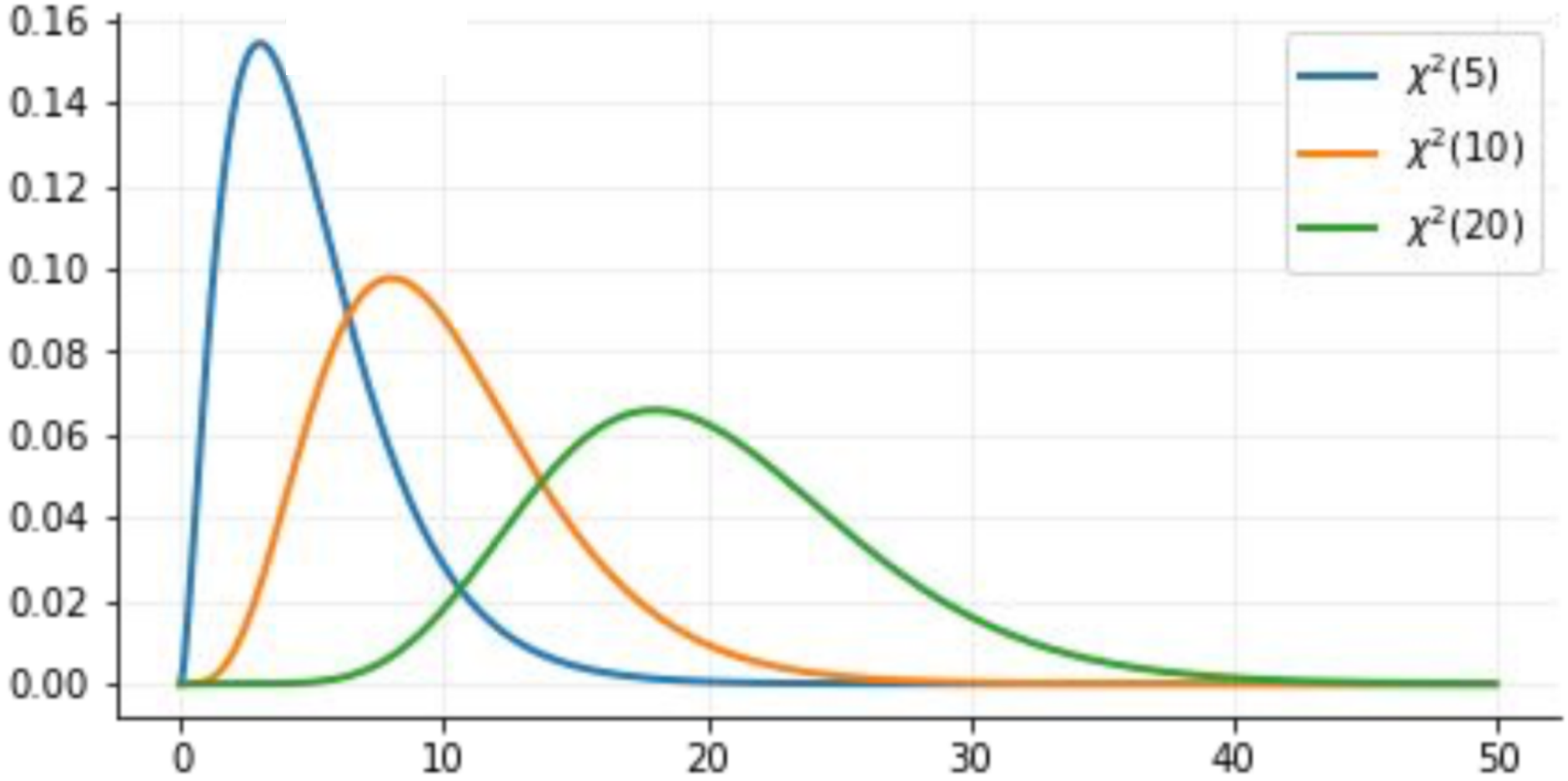
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad .$$

Then the random variable $\frac{(n-1)}{\sigma^2} S^2$ follows the distribution χ^2_{n-1}

Chi^2 distribution

The chi–squared distribution. (χ^2 distribution)

The chi-squared distribution, χ^2_ν is also parameterized by degrees of freedom $\nu = n - 1$



The chi–squared distribution. (χ^2 distribution)

Because the χ^2 distribution is not symmetric, we need to use two different critical values.

α significance level

$$P\left(\chi^2_{1-\frac{\alpha}{2},\nu} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},\nu}\right) = 1-\alpha$$

We are trying to find a CI for σ^2. The $100\cdot(1-\alpha)\%$ CI for σ^2:

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}$$

The two critical values $\chi^2_{1-\frac{\alpha}{2},n-1}$ and $\chi^2_{\frac{\alpha}{2},n-1}$, are chosen to attribute $\frac{\alpha}{2}$ probability to each of the left and right tails.

stats.chi2.ppf(1-α/2, n-1)
stats.chi2.ppf(α/2, n-1)

A confidence interval for the variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52 g. The weight of the packages of candy is known to be normally distributed, but the quality control engineer is concerned that the variation in the produced packages is larger than acceptable. In an attempt to estimate the variance, he selects n=10 bags at random and weighs them. The sample yields a sample variance of 4.2 g^2. Find a 95% CI for the variance, and a 95% CI for the SD.

$n = 10$ $\quad \alpha = 0.05$

$s^2 = 4.2\ g^2$

$\chi^2_{0.025, \nu=9} = \text{stats.chi2.ppf}(.975, 9)$

$= 19.02$

$\chi^2_{1-.025, \nu=9} = \text{stats.chi2.ppf}(.025, 9)$

$= 2.70$

$$\frac{(10-1)\cdot 4.2}{19.02} \leq \sigma^2 \leq \frac{(10-1)4.2}{2.70}$$

$$1.99 \leq \sigma^2 \leq 14$$

CI: $[1.99, 14]$ for σ^2

CI: $[\sqrt{1.99}, \sqrt{14}]$ for σ

Next Time:

- Bootstrapping!