Friday

University of BRISTOL



# Applied Statistics

## Lectures 5 & 6

David Barton & Sabine Hauert

Department of Engineering Mathematics

# Outline

- Functions of random variables
- Central limit theorem
- Sample variance **vs.** population variance
- $\chi^2$ distribution

**OpenIntro Statistics**

Chapter 4, particularly §4.4

# When things aren't normal

CH2M do traffic management for many cities. A common intervention is to use traffic signals to reduce speeding — how do they know it's been successful?

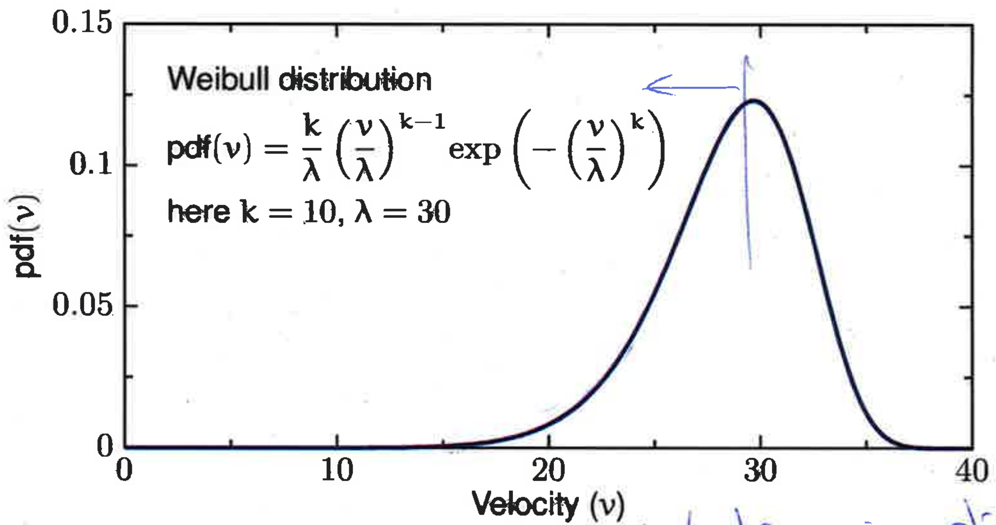
Given the mean and variance of the car velocities before, sample (measure) the car velocities after the change and perform a hypothesis test.

What should the null hypothesis be?

# Distribution of car speeds

Previously used assumption of normally distributed individuals is *probably not a good idea for car velocities!*

The Weibull distribution is probably a better choice (though still not ideal)




distribution $\Rightarrow$ slightly skewed

# Weibull distribution

The Weibull distribution is used in many different areas

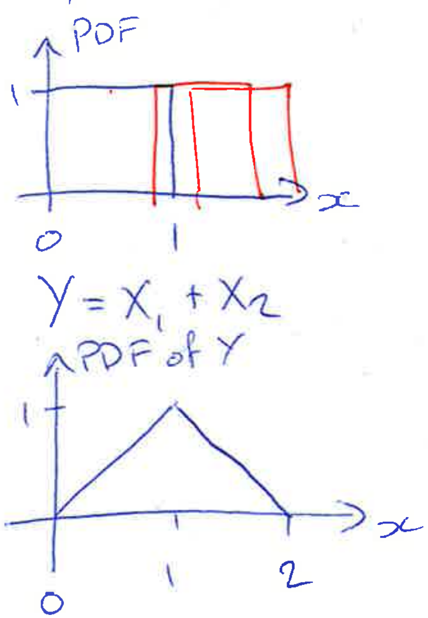
- In reliability engineering and failure analysis.
- In supply chain analysis to represent manufacturing and delivery times.
- In weather forecasting to describe wind speed distributions.
- In communications systems engineering to model fading channels in wireless communications.

*But* there is no known closed form expression for the distribution of the sum of independent Weibull-distributed random variables $X_i$.

So, what is the distribution of the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} X_i?$$

$X_1, X_2 \sim U(0,1)$




# Functions of random variables

In general, it is very hard to determine the distribution of an arbitrary function of random variables — ultimately many of the statistical tests we use are simply because they are the functions we know the distributions for!

In general, when we have two random variables

$$X_1 \text{ with pdf } f(x) \text{ and } X_2 \text{ with pdf } g(x)$$

the distribution of their sum is

$$Y = X_1 + X_2 \text{ with pdf } (f \star g)(x)$$

$X_1, X_2 \sim$ Weibull

where $(f \star g)$ is the convolution integral

$$(f \star g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\,dt$$

(You don't need to be able to do this! Can approximation with Matlab.)

↳ numerically

# Central limit theorem

Fortunately a very useful (but much abused by students!) theorem comes to the rescue.

## Theorem (Central limit theorem)

*Given a random sample of* $n$ *independent and identically distributed (i.i.d.) random variables (individuals)* $\{X_1, \ldots, X_n\}$, *each with mean* $\mu$ *and variance* $\sigma^2$, *the sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} X_i \quad \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

*is normally distributed such that* $\bar{x}\sqrt{n} \sim N(\mu, \sigma^2)$ *as* $n \to \infty$.

Hence for large (but finite) $n$ we have that $\bar{x}$ is approximately distributed as $N(\mu, \sigma^2/n)$. *But it says nothing about the individual* $X_i$ *values!*

if the original distribution is approx symmetric then $n \approx 30$ is good approximation.

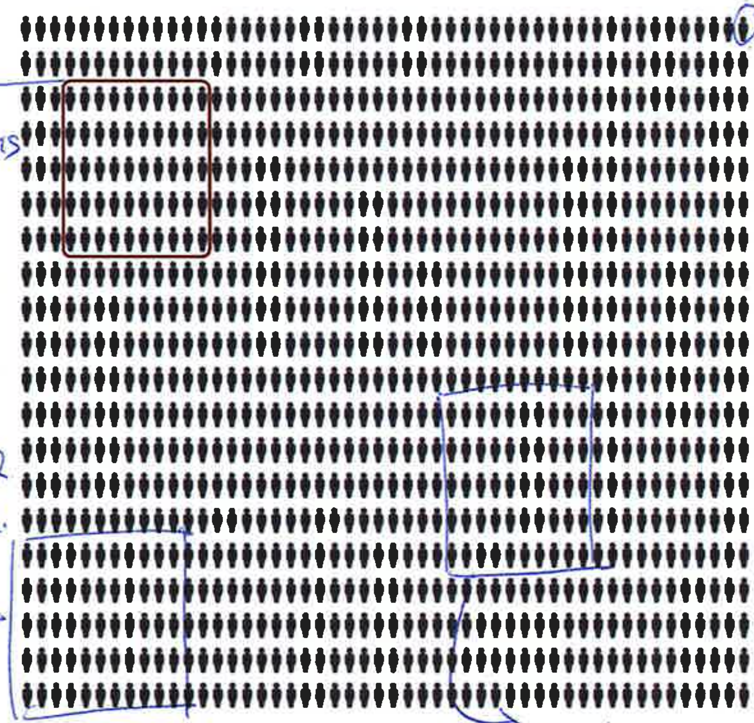
if the orignal distribution is strongly skewed


then $n \gtrsim 100$ is better

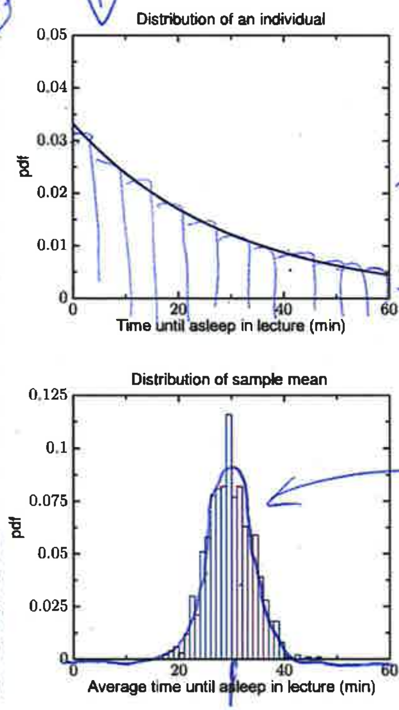
# Distribution of sample mean

50 observations
↳ a single estimate for the mean

calc sample mean.



sample.




exponential distribution.

→ generate histogram of individual observations.

histogram of sample means.

30mins.

# Example from CH2M

How does this help us with the traffic problem? We can now construct the distribution for the null hypothesis.

*Null hypothesis* $H_0$*:* the mean velocity of cars is unchanged after the intervention.

Assume that there are enough individuals in the sample mean for the central limit theorem to apply (approximately). Thus from the parameters of the original Weibull distribution we have that

variance of the original Weibull distribution.

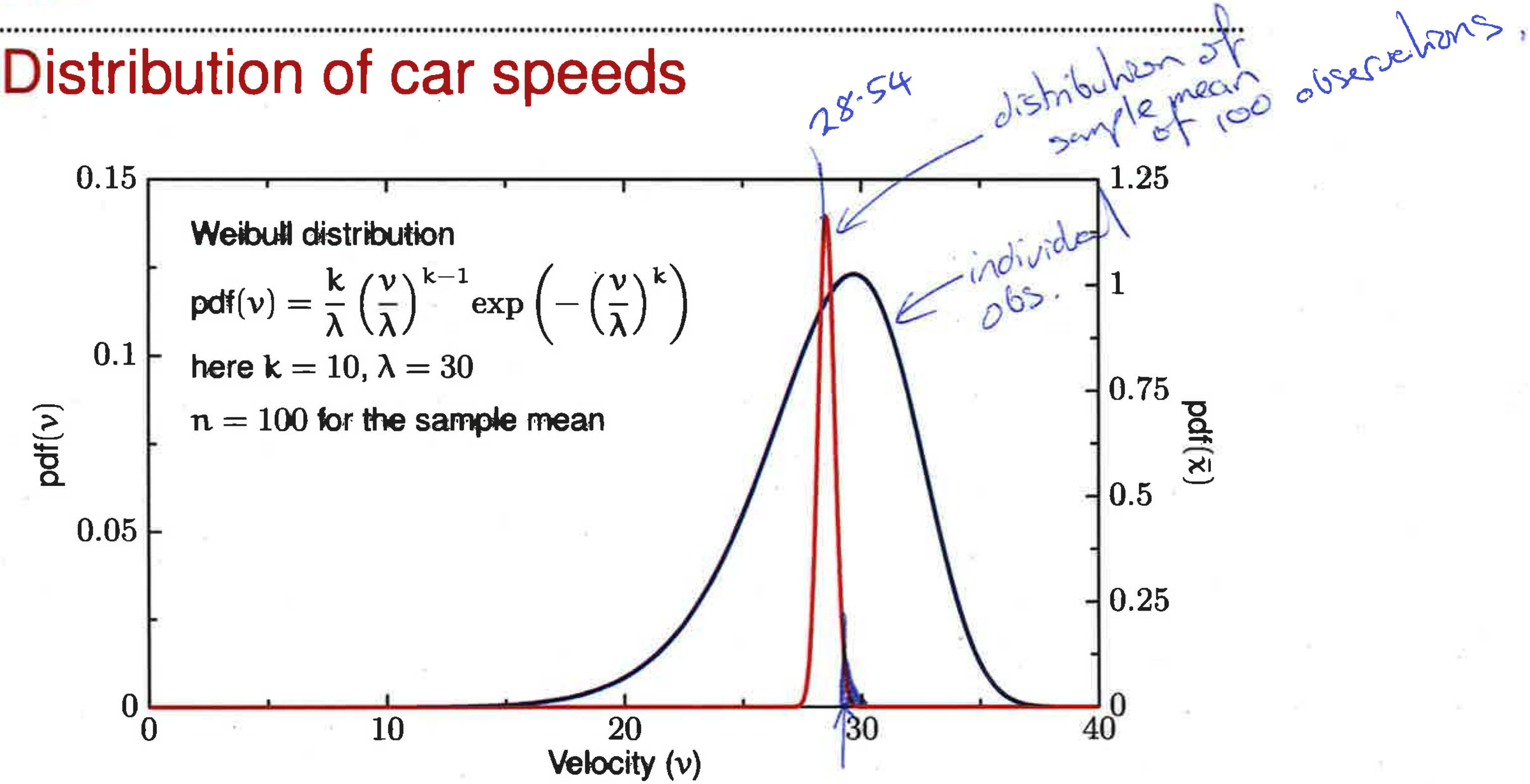
$$\bar{x} \sim N\left(\underline{28.54}, \frac{11.79}{n}\right)$$

mean

[For the Weibull distribution we have that $\mu = \lambda\Gamma\left(1+\frac{1}{k}\right) = 28.54$ and $\sigma^2 = \lambda^2\left(\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right) = 11.79$. (Formulae that you look up rather than memorise...)]

Gamma function.

# Distribution of car speeds



# Test the hypothesis

After the intervention has been implemented the sample mean velocity of 100 cars is found to be 29.5 mph. To 5% significance, can we reject the null hypothesis?

5% significance means that if the probability of seeing the effect we see *or something more extreme* is less than 5% then we reject the null hypothesis and accept the alternative hypothesis $H_1$.

In this case, that the mean value has been changed by the intervention.

# Test the hypothesis — p-values

The difference between the measured and the hypothesised means is $29.5 - 28.54 = 0.96$. Hence we have

$$p = 2\min\left(P(\bar{x} \leq 29.5),\ \underbrace{P(\bar{x} \geq 29.5)}\right)$$

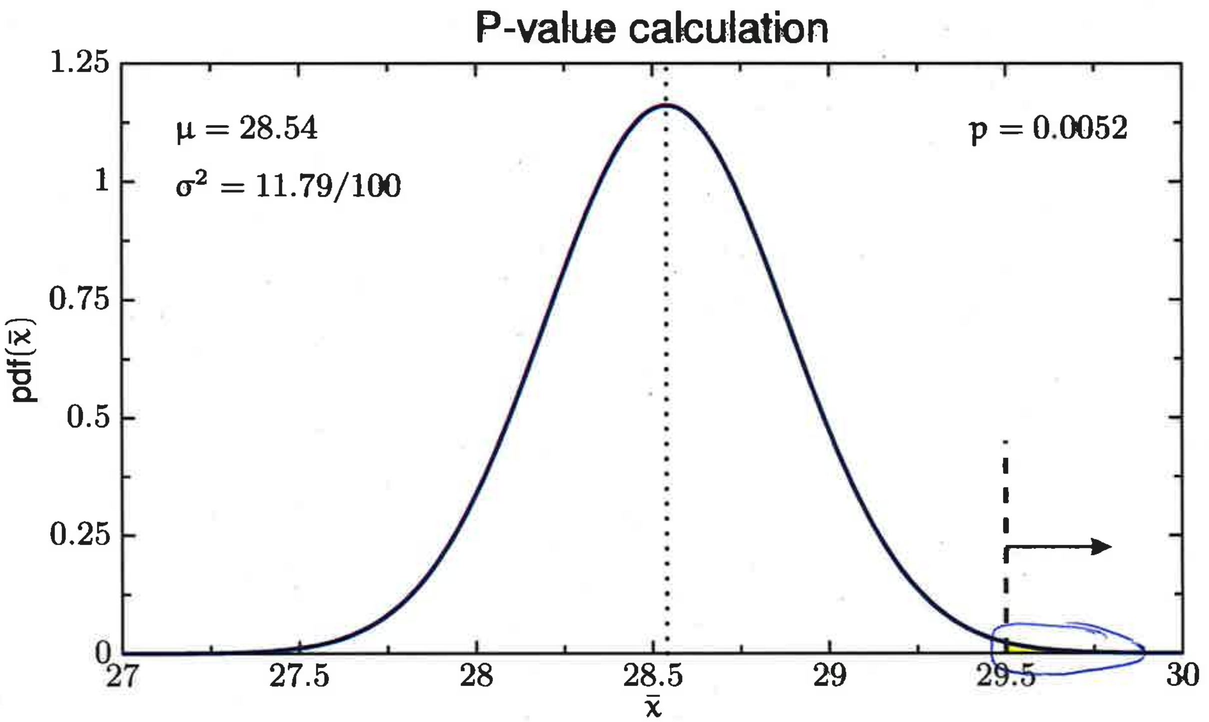
$$= 2P(\bar{x} \geq 29.5) \qquad \bar{x} \sim N\left(28.54, \frac{11.79}{100}\right)$$

$$= 2P\left(z \geq \frac{29.5 - 28.54}{\sqrt{11.79/100}}\right) \qquad z \sim N(0,1)$$

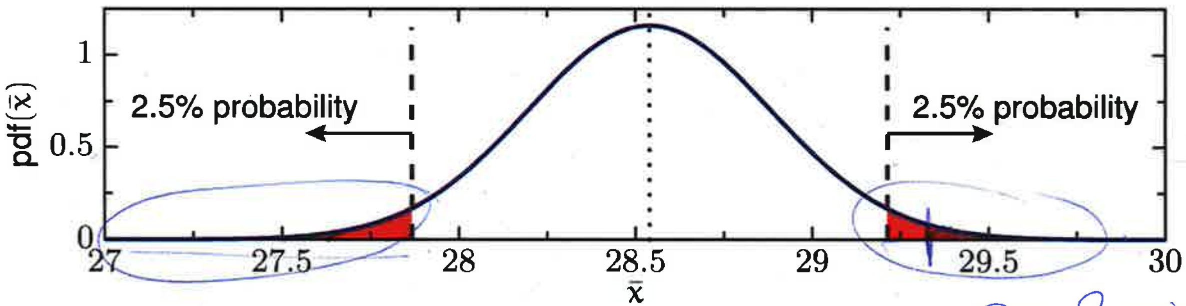
$$= 2P(z \geq 2.80) = 0.0052 < \text{significance level } (5\%)$$

Thus we reject the null hypothesis

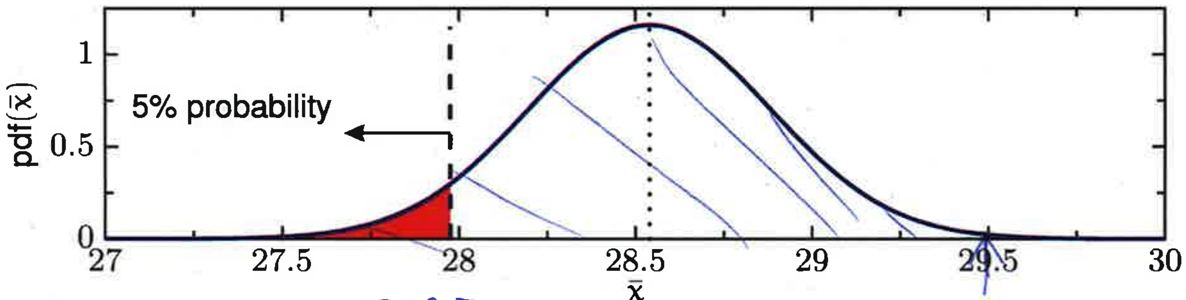
# P-value graphically



# Back to the traffic — one tail or two?




$P = 2\min(P(\bar{x} \leq 29.5), P(\bar{x} \geq 29.5))$



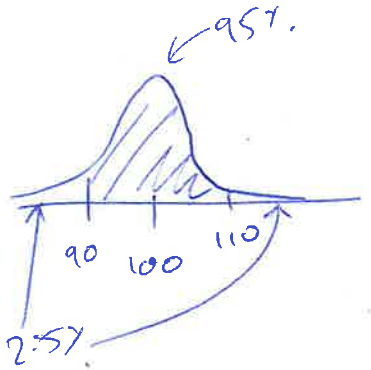

$P = P(\bar{x} \leq 29.5) \approx 1$

# Exercises

The underlying distribution for each observation is normal. $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

An electronics manufacturer produces capacitors with a (mean) rated capacity of 100 nF and a standard deviation of 3 nF.

1. If I buy 4 and measure their capacities as 95, 101, 91, 97 nF should I believe the manufacturer (use 5% significance)? $\rightarrow$ No don't believe manufacturer.
2. What about if the manufacturer instead says that 95% of the capacitors should lie within ±10 nF of the rated capacity?




$P(X < 90) = 0.025$

$P(z < \frac{90-100}{\sigma}) = 0.025$

$\frac{90-100}{\sigma} = -1.96.$

$\Rightarrow \sigma = \frac{10}{1.96} = 5.1$

$H_0$: the manufacturer is telling the truth

$X \sim N(100, 3^2)$ $\quad \bar{x} \sim N(100, 3^2/4)$

5% two-tailed test.

$\bar{x} = 96\,nF$

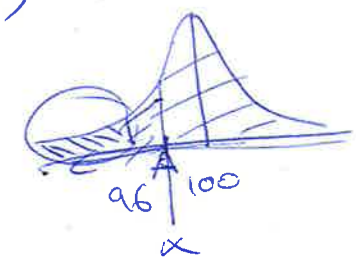
$P = 2\min(P(\bar{x} \leq 96), P(\bar{x} \geq 96))$

$= 2P(\bar{x} \leq 96) = 2P(z \leq \frac{96-100}{3/2})$

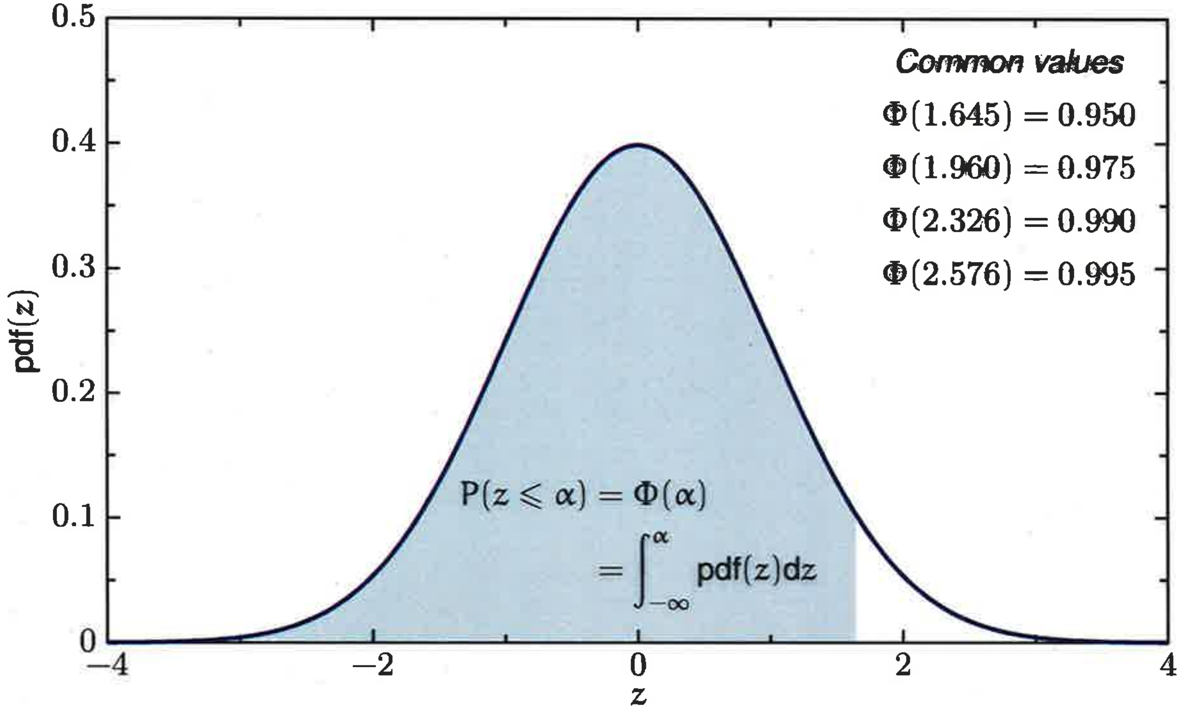
$= 2P(z \leq -8/3)$

$-2.666$ less than critical value $\alpha = -1.96$ $\rightarrow$ reject $H_0$.

$P(z \leq \alpha) = 0.025$



# Standard normal distribution



# Monte Carlo simulation as an alternative

```
% 10,000 groups of 20 observations, lambda=30, k=10
x = wblrnd(30, 10, 20, 10000);
% mean of each of the groups of 20
y = mean(x);
% plot empirical PDF
histogram(y, 'Normalization', 'pdf')
% sorted list of random values
y2 = sort(y)
% lower bound at 2.5%
y2(length(y)*0.025)
% upper bound at 97.5%
y2(length(y)*0.975)
```

# Test statistics — sample mean

Previously we've considered the sample mean as a test statistic where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

Notice how parameters from the hypothesis ($\mu$ and $\sigma$) appear in the distribution. It's better (read: more standard, less error prone) for all the parameters to appear in the test statistic only.

## Definition (Test statistic for the sample mean (normal))

The test statistic for the sample mean of a normally distributed random sample with *known variance* is

pop mean

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

pop. variance.

# Test statistics — standard error

Note that the test statistic takes the general form

$$\frac{\bar{x} - \mu}{SE}$$

sample mean is an estimate for the population mean.

where $SE$ is the standard error

## Standard error

The standard deviation associated with an estimate is called the *standard error*. It describes the typical error or uncertainty associated with the estimate.

For the estimate of the sample mean, the standard error is

$$SE = \frac{\sigma}{\sqrt{n}}$$

# Test statistics — sample mean

Notice that the sample mean test statistic is calculated with the population variance! (In the case of hypothesis testing, it means that the variance is stated as part of the hypothesis.)

$\bar{x}$ is normally distributed (either relying on the samples being normal, or the central limit theorem when there are a large number of samples) and so the test statistic is just a normal variable multiplied by some constants giving another normal variable.

What happens when you don't know the variance a priori? Estimate it.

Sample mean → population mean
Sample variance → population variance
Random variables ↳ with own distributions
↳ constants.



# Sample variance

As well as estimating the sample mean from the data, we can also estimate the sample variance.

## Definition (Sample variance)

$\frac{1}{n-1}$ → use this!

The variance of a sample is defined as

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2} = \left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - \bar{x}^2.$$

This is a *biased* estimation of the population (or true) variance $\sigma^2$!

$$E\left[s^2\right] = \boxed{\frac{n-1}{n}}\sigma^2 \neq \sigma^2$$

where $E[\cdot]$ is the expected value.
(Note: $E\left[s^2\right] \to \sigma^2$ as $n \to \infty$ so the estimate is *consistent*.)

# Population variance

To get an *unbiased* estimate of the population (or true) variance $\sigma^2$ we must use the correction factor $[n/(n-1)]s^2$. Hence

Bessel's correction.

$$E\left[\frac{n}{n-1}s^2\right] = \sigma^2$$

Consequently, sometimes the expression

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2}$$

use this!

is used directly for the sample mean to avoid this problem.

*Be clear which you are using!*

Why does this happen? It's because $\bar{x}$ is used in the equation for the sample mean rather than $\mu$. If $\mu$ is used, the $1/n$ factor is correct.

# Sample variance as a random variable

Like the sample mean, since the sample variance is a function of random inputs (the $X_i$) it is a random variable with a particular distribution.

The sample variance of $n$ samples follows the $\chi^2$ (chi-squared) distribution with $n - 1$ degrees of freedom.

## Definition (Test statistic for the sample variance (normal))

The test statistic for the sample variance of a normally distributed random sample with $n$ samples is

$$\boxed{(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}}$$

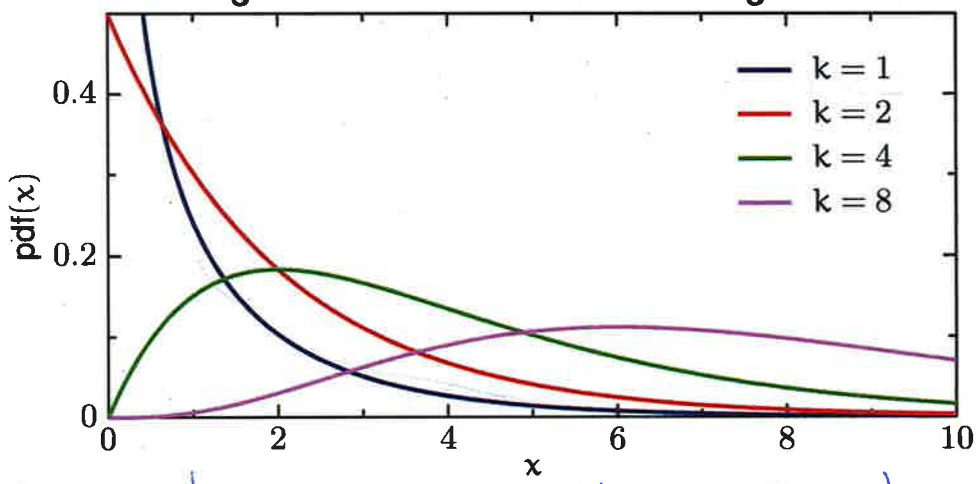
This can be used to test a hypothesis on the variance of a sample (but not practically useful due to high sensitivity to non-normal data).

# $\chi^2$ distribution

The $\chi^2$ distribution is defined for positive values by the PDF

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

k is the number of degrees of freedom and $\Gamma$ is the gamma function.




Sample variance is always positive.
$\Rightarrow \chi^2$ is always positive.

# $\chi^2$ distribution — properties

The key fact that links $\chi^2$ to the normal distribution is that for $X_i \sim N(0,1)$ we have that

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2.$$

I.e., the sum of squared normally distributed random variables is $\chi^2$ distributed with $n$ degrees of freedom.

Also, it “inherits” some nice properties from the normal distribution; if

$$X_1 \sim \chi_{k_1}^2 \quad \text{and} \quad X_2 \sim \chi_{k_2}^2$$

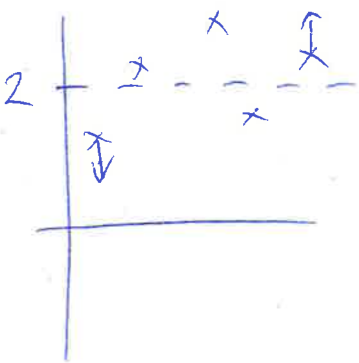
then

$$X_1 + X_2 \sim \chi_{k_1+k_2}^2,$$

that is, the degrees of freedom add together.

# Degrees of freedom

mean of 2



The degrees of freedom are the number of data points minus the number of constraints you've placed on the data. In the absence of constraints, $n$ measurements (data points) = $n$ degrees of freedom.

The test statistic for the sample variance

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

uses $n-1$ degrees of freedom since in calculating the test statistic the value of $\bar{x}$ is fixed (constrained) — there is an $n-1$ dimensional set of data points that all give the same sample variance.

*Geometric example:* consider the point $(x, y, z)$ in 3D space. If I add the constraint that $x^2 + y^2 + z^2 = 1$, I've constrained the point onto a 2D surface in 3D space, and so the number of degrees of freedom have been reduced.

# Quote of the day

## Ronald Fisher

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.


### Exercises

- 4.1–4.4, 4.23–4.24, 4.27–4.30 from OpenIntro Statistics