1 Distributions

Distributions

Words

Given a distribution X, the

probability distribution function (pdf) (continuous) or **probability mass function (pmf)** is the probability that the variate has value x:

$$\Pr(a \leq X \leq b)$$

or

$$\Pr(X = a)$$

cumulative probability function (cdf) is the probability that the variate is less than x:

$$\Pr(X \leq x) = \int_{-\infty}^{x} \text{pdf}(x)\,\mathrm{d}x \text{ or } \sum_{i \leq x} \Pr(X = x)$$

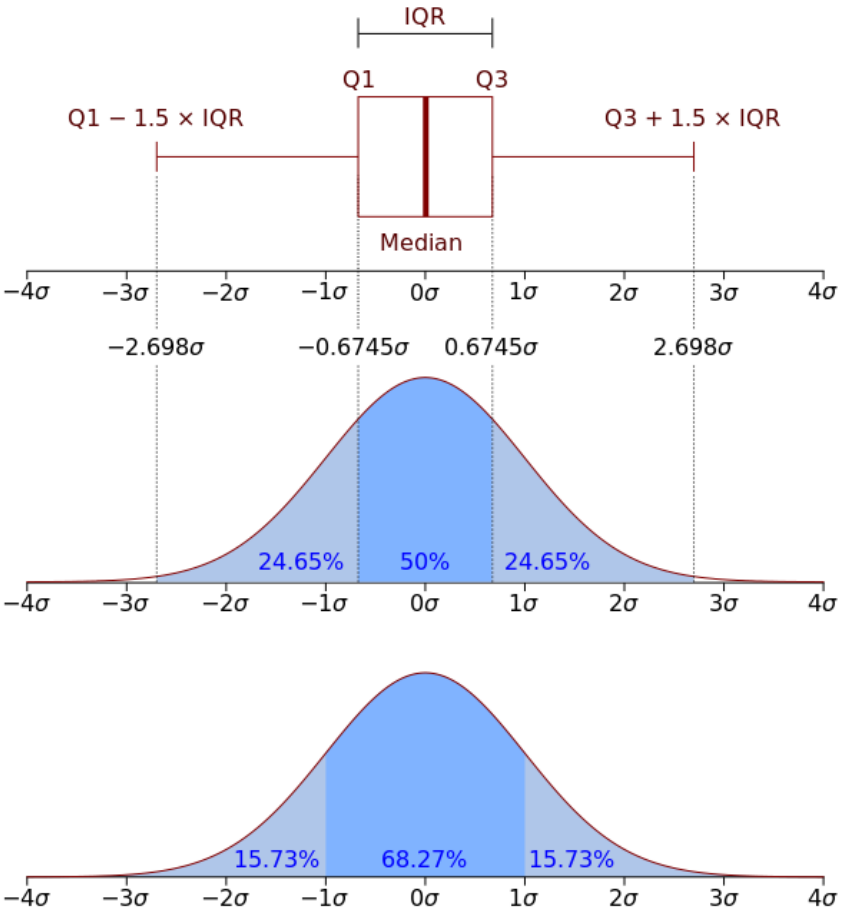
percent point function (ppf) is the inverse of the cdf. Given a probability, what's x? Also called the **inverse distribution**.

survival function (sf) is the probability that the variate takes a value greater than x:

$$\text{ss}(x) = \Pr(X > x) = 1 - \text{cdf}(x)$$

inverse survival function (isf) is the inverse of the survival function:

$$\text{isf}(\alpha) = \text{ppf}(1 - \alpha)$$



Code lab

ipython notebook

Uniform distribution

Takes value 1 with probability 1.

Model	identical events
Parameters	none
Mean	$\frac{1}{2}$
Variance	$\frac{1}{12}$

Does this mean I have probability 1 of picking any number on the unit interval?

http://docs.scipy.org/doc/scipy/reference/stats.html\#continuous-distributions

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.uniform.html\#scipy.stats.uniform

Bernoulli distribution

Takes value 1 with probability p and 0 with probability $1 - p$.

Model	turning coins
Parameters	$p \in [0, 1]$
Mean	p
Variance	$p(1 - p)$

The pmf(k) (mass function) is $1 - p$ if $k = 0$ and p if $k = 1$.

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bernoulli.html\#scipy.stats.bernoulli

Geometric distribution

Two definitions:

- The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, \dots\}$
- The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $\{0, 1, 2, 3, \dots\}$

Wikipedia, geometric distribution

Model	...
Parameters	$p \in (0, 1]$
Mean	$\frac{1}{p}$ or $\frac{1-p}{p}$
Variance	$\frac{1-p}{p^2}$

This computes the version with $k - 1$ failures and then a success.

No canonical example of model. Attempts to succeed at a task, for example. Time to die from your daily commute.

Does this model Russian roulette? (No, not independent.)

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.geom.html\#scipy.stats.geom

Poisson distribution

Probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Model	radioactive decay, network packets
Parameters	$\lambda \in \mathbb{R}^+$
Mean	λ
Variance	λ

$$\text{pmf}(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

(Sometimes we say μ instead of λ.)

Examples: Deaths in Greater London, λ is the mean rate of deaths per day.

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.poisson.html\#scipy.stats.poisson

Binomial distribution

$\mathbf{B}(n, p) =$ Number of successes in a sequence of n independent bernoulli trials (yes/no experiments), each of which yields success with probability p.

Model	sequences of coin tosses
Parameters	n, p
Mean	np
Variance	$np(1-p)$

$$\text{pmf}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in \{0, 1, \ldots, n\}$.

Often convenient to use the normal distribution as an approximation.

Binomial distribution: Example

19% of the British public smokes. A study of 500 people reports 90 smokers. What is the probability of finding 90 or fewer smokers in a sample of 500 UK residents chosen at random?

http://en.wikipedia.org/wiki/Smoking_in_the_United_Kingdom

Exercise

TODO: Make notebook

Show mathematically, then with scipy.

```
n=500, p=.19
```

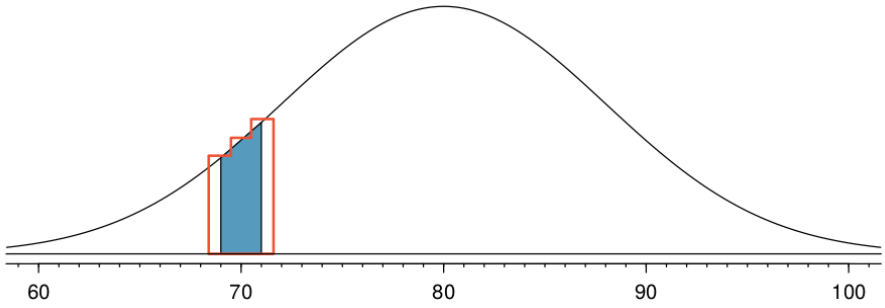
```
ss.binom.cdf(90, n, p)
```

The normal approximation is reasonable, since $(.2)(500) = 100 > 10$ and $(1-.2)(500) = 400 > 10$.

Use $\mu = np = 100$, $\sigma = \sqrt{np(1-p)} = \sqrt{80} \approx 8.9$, $\mathcal{N}(100, 8.9)$. Then $Z = \frac{59-100}{8.9} \approx -4.6$.

`mu = n*p, sigma = math.sqrt(n*p*(1-p)) ss.norm.cdf(90, mu, sigma)`

We'll talk about Z value soon.



Rule of thumb: The binomial distribution is reasonably approximated by the normal distribution when both np and $n(1-p)$ are greater than about 10. Also a sufficient range (i.e., not just a couple values). Can mitigate by decreasing lower cutoff by 0.5 and increasing upper cutoff by 0.5.

Negative binomial distribution

How many Bernoulli trials with parameter p until we have n successes?

Model	
Parameters	p, n
Mean	$\frac{pn}{1-p}$
Variance	$\frac{pn}{(1-p)^2}$

$$\text{pmf}(k) = \binom{k+n-1}{n-1} p^n (1-p)^k$$

for $k \geq 0$.

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.nbinom.html\#scipy.stats.nbinom

Comparing discrete distributions

Binomial distribution. Fixed number of trials, measures probability of success.

Negative binomial distribution. Fixed number of successes, measures probable number of trials.

Poisson distribution. Fixed number of trials, measures probable number of successes.

Normal distribution

$\mathcal{N}(\mu, \sigma^2)$, about which we will say a great deal over the next hour and the rest of the day.

Model	cf. CLT
Parameters	μ, σ^2
Mean	μ
Variance	σ^2

$$\text{pdf}(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Effects of varying parameters (do

import matplotlib.pyplot as plt import numpy as np import matplotlib.mlab as mlab import math

mean = 0 variance = 1 sigma = math.sqrt(variance) x = np.linspace(-3,3,100) plt.plot(x,mlab.normpdf(x,mean,sigma))

plt.show()

import numpy as np import matplotlib.pyplot as plt from scipy.stats import norm

Plot between -10 and 10 with .001 steps. range = np.arange(-10, 10, 0.001) # Mean = 0, SD = 2. plt.plot(range, norm.pdf(range,0,2))

Bell curve. Carl Friedrich Gauss.

Note that unimodal and roughly symmetric does not necessarily mean normal.

Nonetheless, normal is often a good approximation.

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html

Z score

$$Z = \frac{x - \mu}{\sigma}$$

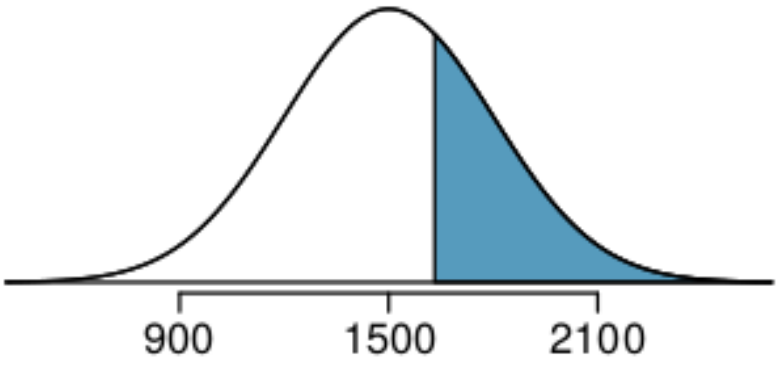
We compute the Z score for each observation.

Z scores are a coordinate transform to $\mathcal{N}(0, 1)$.

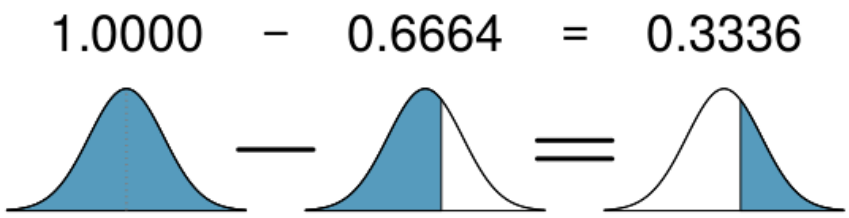
So we can use to compute percentiles.

Example

The scores on an exam are approximately $\mathcal{N}(1500, 300)$. What is the probability a random exam taker (one about whom we know nothing a priori) scores above 1630?



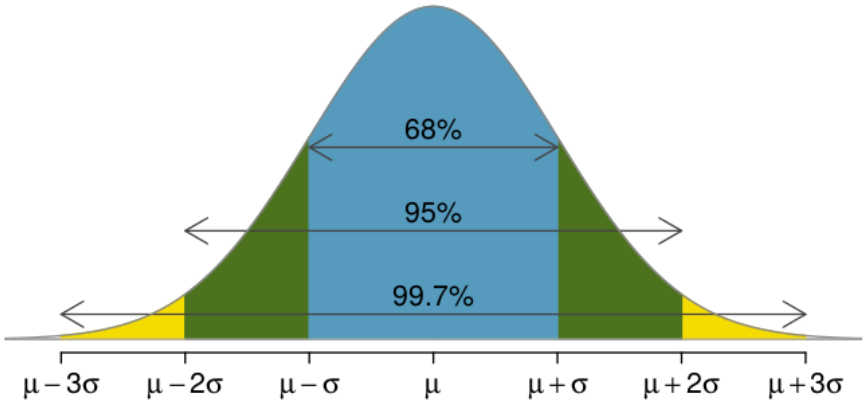
$$Z = \frac{1630 - 1500}{300} = .43$$



Exercise: scipy.

We can also go backwards, from percentile to Z score to values.

65-95-99.7 Rule

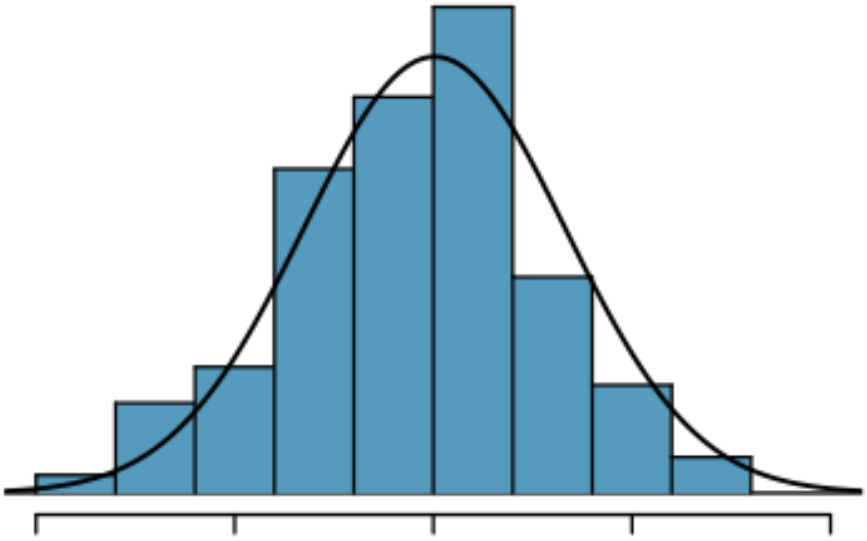


Exercise: plot this.

Exercise: Demo with scipy probabilities of falling outside $n\sigma$ for $n = 1, \ldots, 7$.

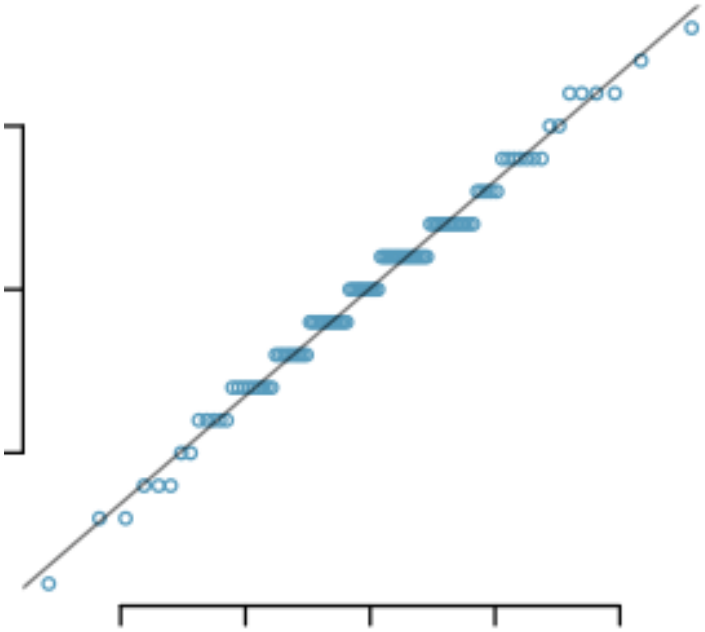
Evaluating Normal Approximations

Easy technique 1: visually compare to normal plot.



Demo this in matplotlib.

Easy technique 2: normal probability plot.



Also known as a quantile-quantile plot.

Demo these in matplotlib, both prepared distributions (normal and not) and with norm() and different values of n.

2 Inference

Inference

Inference

This is the important part.

Goal: Understand the quality of parameter estimates.

Examples:

- How close is $\overline{x}$ to μ?

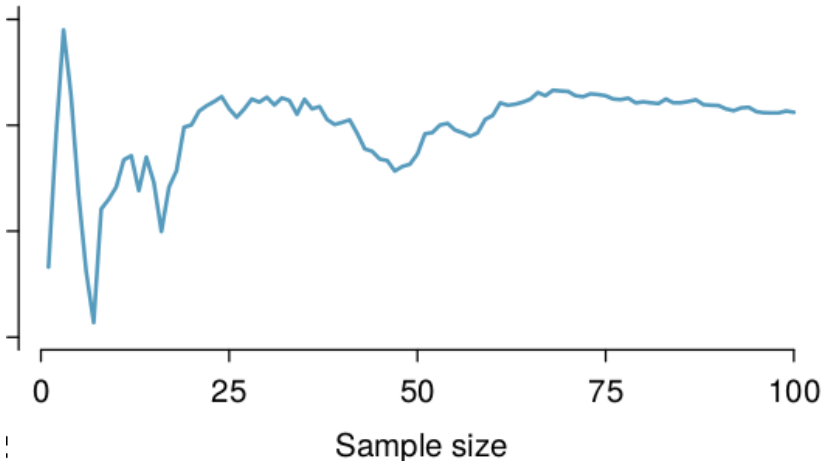
Point estimates

Population mean (μ) $\neq$ sample mean ($\overline{x}$).

We call $\overline{x}$ the *point estimator* of the population mean μ, a parameter of the distribution.

Point estimate: if you have to guess, this is it.

Running mean. Sequence of partial sums (divided by number in sum).



Exercise!

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

"All problems in computer science can be solved by another level of indirection" —*David Wheeler* "All problems in computer science can be solved by another level of indirection, except of course for the problem of too many indirections." —*David Wheeler*

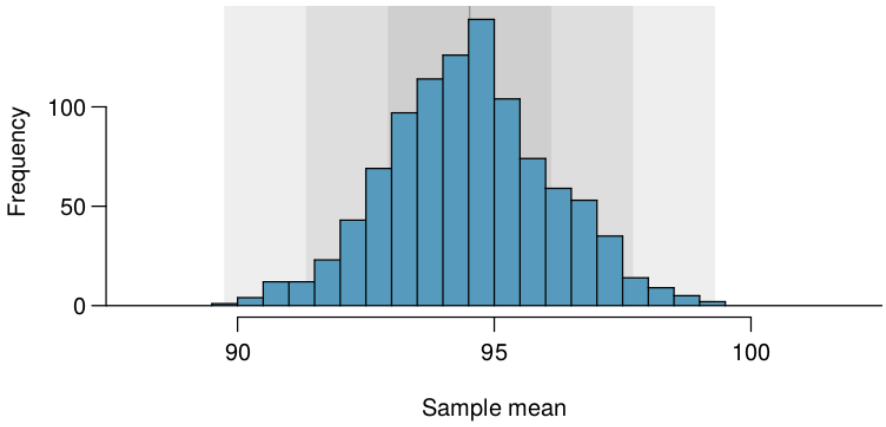
*http://en.wikipedia.org/wiki/David_Wheeler_\%28British_computer_*

Smaller sampling variation means probably more accurate value.

Think of sampling distribution as the set of subsets from which all samples come.

Understanding sampling distributions is central to understanding inference.

Sampling distribution



- Sampling mean is unimodal and approximately symmetric.
- It is centred at population mean.
- The standard deviation of the sample mean tells us how far a point sample's mean is likely to be from the population mean. In other words, how much error we are likely to have in the point estimate's mean. **Standard error.**

Exercise: Generate uniform population, sample, and plot sampling distribution.

Exercise: Generate highly skewed population, sample, and plot sampling distribution.

In real life, we don't have access to the population parameters. We have to *estimate* them from samples. So we can't *know* the standard error.

Easy exercise:

- Would you rather have a small sample or a large sample when estimating a parameter?
- Would you expect a point sample based on a small sample to have smaller or larger standard error than a point sample based on a large sample?

Given n observations, the standard error of the sample means is

$$\text{S.E.} = \frac{\sigma}{\sqrt{n}}$$

Cf. $\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y)$

Usually use $\overline{s}$ instead of σ.

Rules of Thumb

When is the preceding a good approximation?

- $n > 30$
- The population is not too skewed.

When is the sample likely to be independent?

- Use simple random sampling.
- $n < 10\%$ of the population.

Summary

- Point estimates estimate population parameters.
- Point estimates have error and vary among samples.
- The standard error quantifies the variation among point estimates.

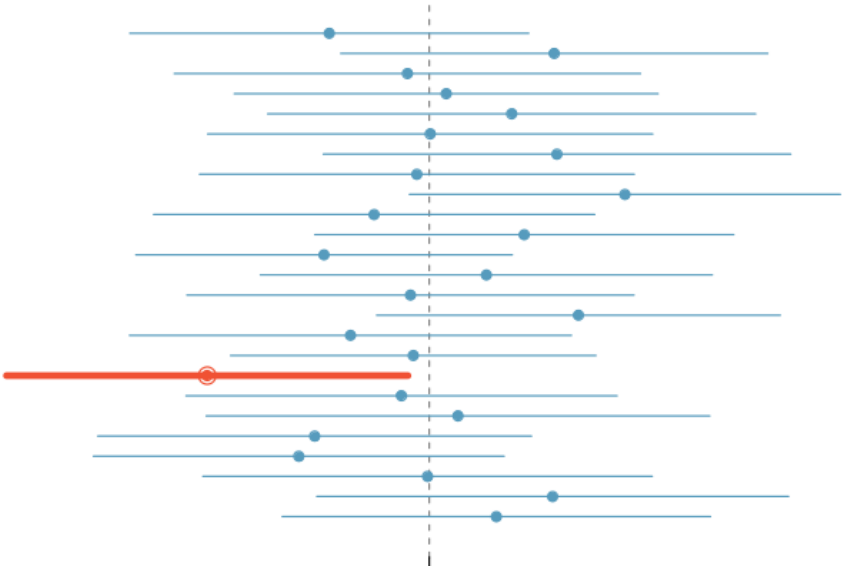
Confidence intervals

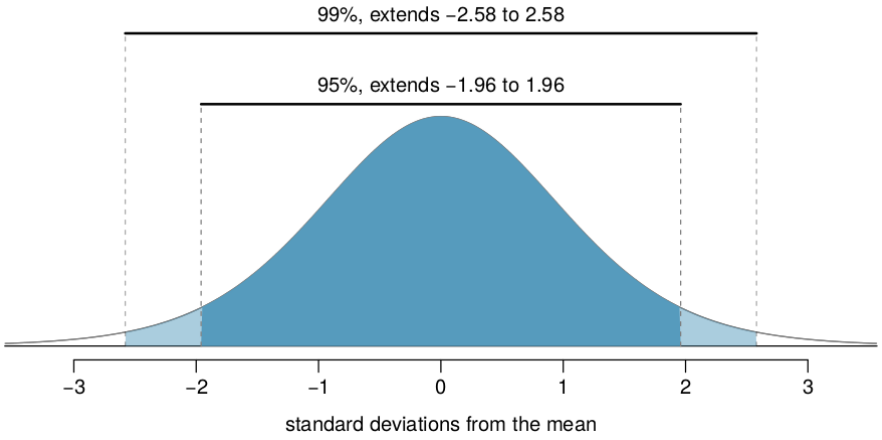
Sample n points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

95% confidence means ± 2 S.E.

95% confidence means ± 1.96 S.E.





Sampling is usually expensive.

Reminder: Independent random samples!

We call z^* S.E. the *margin of error*.

Correct language: "We are 95% confident that the population parameter is between..."

Incorrect language: describe the confidence interval as capturing the population parameter with a certain probability.

This is one of the most common errors: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.

Another especially important consideration of confidence intervals is that they only try to capture the population parameter. Our intervals say *nothing* about the confidence of

- capturing individual observations
- a proportion of the observations
- about capturing point estimates

Confidence intervals only attempt to capture population parameters.

p-value

Do people die later now than 10 years ago?

Compute the mean death age from 10 years ago. Compute a sample mean now and confidence interval.

H_0: no H_A: yes

If the mean is within the confidence interval, we reject H_A. If the mean is outside the confidence interval, we accept H_A.

H_0 should usually be an equality.

But maybe the mean falls near the edge, or way outside. How to hint?

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative.

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.

The p-value is a conditional probability.

It answers "what is the probability of the observed value if H_0 is true, assuming a normal sample distribution?"

So it is the area under a bit of the tail of a Gaussian.

Example: Study shows that 10 out of 100 smokers will develop a certain cancer. *Bad conclusion:* For a given smoker, there's a 10% chance he'll develop this cancer. *Good conclusion:* If I take a large enough sample of smokers, about 10% of them will develop this cancer.

Example: p-value

Googlers gain on average 7 kgs within one year of working at the company. *(This may not be true, but it's a common joke.)*

Facebook thinks their food is better, they say their employees gain more weight in the first year. *This would be a perverse argument for joining facebook.*

H_0: $\mu = 7$ H_A: $\mu > 7$

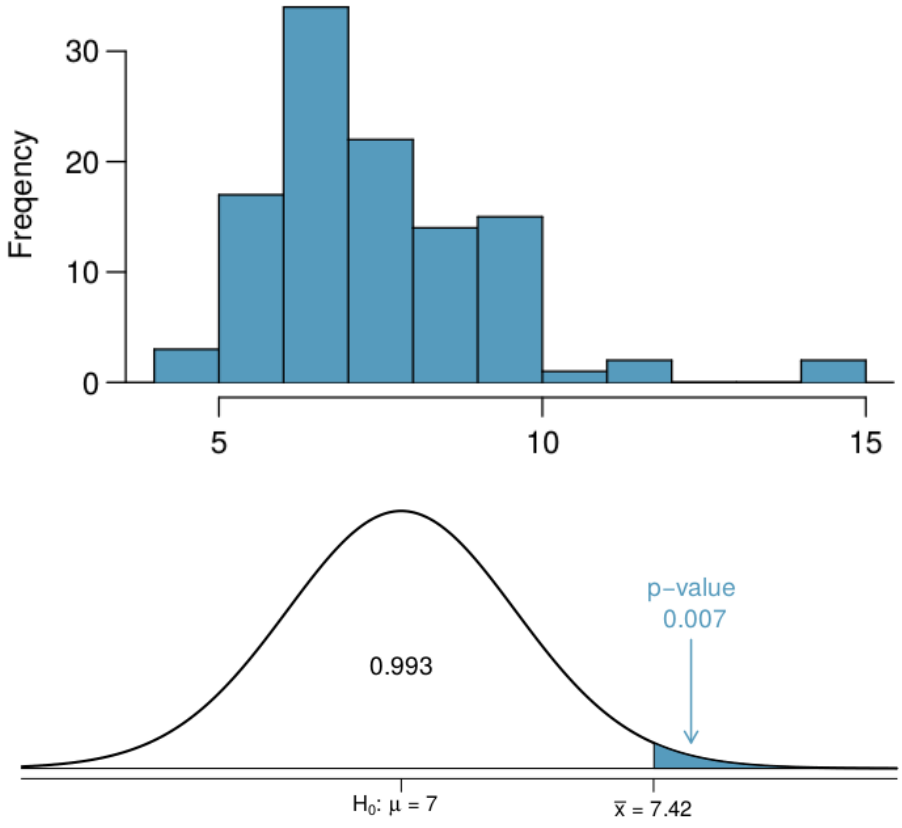
Sample $n = 110$ facebookers, $\overline{x} = 7.42$ and $s = 1.75$.

S.E. $= \frac{s}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} \approx 0.17$

$\mu > 7$ is a one-sided hypothesis. Use two-sided usually. Decide before data collection.

Check that normal approximation is ok:

- simple random sample, $< 10\%$, so independent
- sample size > 30
- moderate skew, outliers not too extreme



Shaded tail represents chance of observing $\overline{x}$ conditional on H_0 being true.

That is, it's the p-value.

$$Z = \frac{\overline{x} - \mu}{\text{S.E.}} = \frac{7.42 - 7}{0.17} \approx 2.47$$

$$\Pr(\overline{x} < 7.42) = .993, \text{ so } p = 1 - .993 = .007.$$

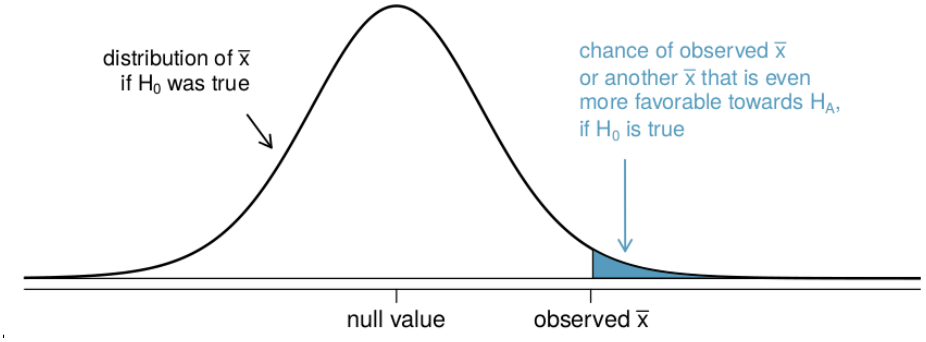
So reject H_0.

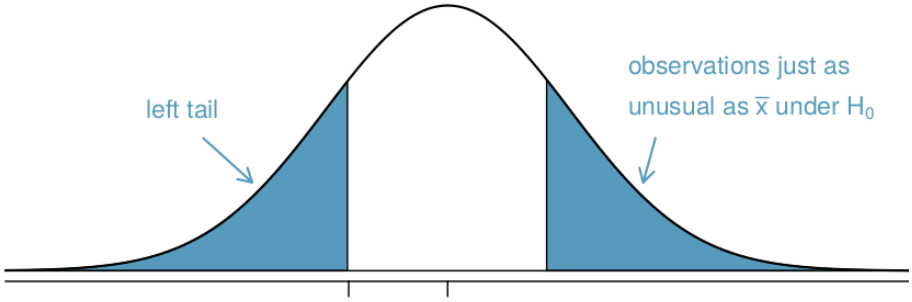
Note that $p < .05$ is very arbitrary and now often unacceptable in science.

What is the cost of a Type I error (incorrectly rejecting a true H_0)?

What is the cost of a Type II error (failure to reject a false H_0)?

Useful to restate hypothesis in plain language so people understand.





Central Limit Theorem (CLT)

The distribution of $\overline{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

Open Statistics

Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

http://en.wikipedia.org/wiki/Central_limit_theorem

Suppose $\{X_1, X_2, \dots\}$ is a sequence of i.i.d. random variables with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} N(0,\ \sigma^2).$$

http://en.wikipedia.org/wiki/Central_limit_theorem

Vocabulary

A point estimate is **unbiased** if the sampling distribution of the estimate is centred at the parameter it estimates.

An unbiased point estimate does not significantly over- or under-estimate the parameter.

Significance (revisited)

- Flip a coin N times. Get 49% heads. Is the coin fair?
- Congress/parliament has N members of whom m are male. Do we discriminate against women?[5mm]
- Careful about what we conclude. Here we can't conclude anything about cause or source.

Code lab: A/B testing

Exercises:

For each,

- What are H_0 and H_A?
- Should we adopt method B based on the evidence?

A/B Jeopardy

a game in pairs

Questions?

purple.com/talk-feedback

3 Break

Break