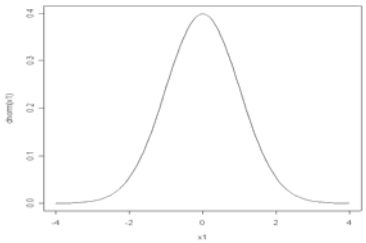
Statistics and Probability

Continuous RVs (Normal); Confidence Intervals



http://www.isrec.isb-sib.ch/~darlene/geneve/

Outline

- Continuous random variables
- Normal distribution
- CLT
- Point estimation
- Confidence intervals

Continuous distributions

- Not all RVs are discrete…
  - Temperature at a certain time and place
  - Height of a randomly chosen person
  - Fluorescence intensity at a spot on a microarray
  - *etc.*

Density function for continuous RV

- The *density function* for a continuous RV X does ***NOT*** have the same interpretation as in the discrete case; in particular, it is ***NOT*** $P(X = x)$
- For a continuous RV, any *particular value has probability 0* of occurring
- Instead, we interpret the density as the height of the 'histogram' for the RV (called the '*density curve*')
- The total area under the density curve = 1

Distribution function for a RV (review)

- The *(cumulative) distribution function* (cdf) for (any) RV X is

$$F(x) = P(X \leq x)$$

- The cdf satisfies:
  1. $F(x)$ is *nondecreasing* for all x
  2. $F(-\infty) = 0$
  3. $F(\infty) = 1$

Expectation of a continuous RV

- The *expected value* of a continuous RV X with density $f(x)$ is

$$E[X] = \int_x x*f(x)dx$$

- This integral is just the continuous analogue of summation in the discrete case

Standard units

- Standard units (SUs), also sometimes called *z-scores*, tell how many SDs above or below the mean (average) a particular observation is
- To convert a value x into standard units z, subtract the mean from the value, then divide that result by the SD:

$$z = (x - \text{mean})/\text{SD}$$

- Subtracting the average from each variable value x has the effect of making the average of the z's be 0; dividing by the SD makes the SD of the z's be 1.

Why standard units?

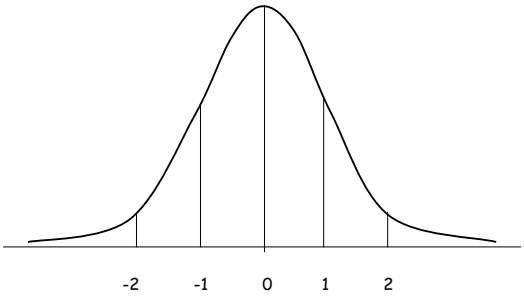
- For *comparing* two (or more) sets of data, it is often useful that values be expressed in the same units
- Detection of suspected *outliers* is often carried out in terms of standard units
- Standard units are important for using the *normal distribution*

Normal distribution

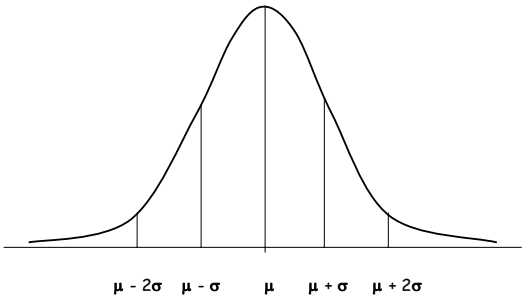
- The *histogram* for the normal distribution looks like a (symmetric) 'bell-shaped' curve

- For the *standard normal* distribution, the mean is 0 and the SD is 1
- Concerning the *AREA under the curve*, about
  - 68% is within 1 SD of the mean
  - 95% is within 2 SDs
  - 99.7% is within 3 SDs

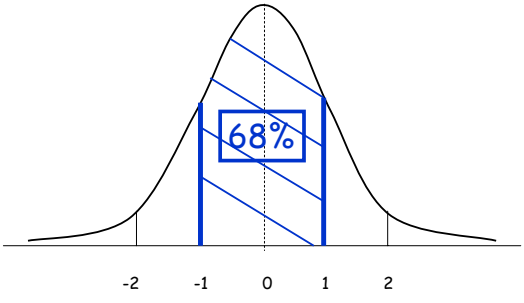
Standard normal distribution



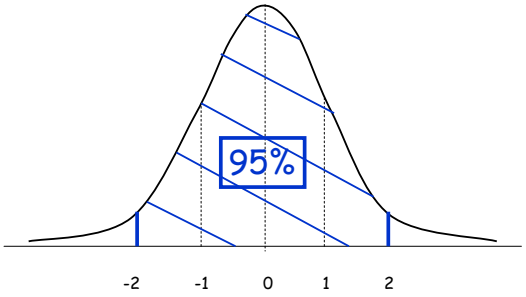
General normal distribution



Within 1 SD



Within 2* SDs (* really 1.96)



Importance of the normal distribution in statistics

- Convenient mathematical properties
- Variations in a number of physical experiments are often approximately normally distributed
- *Central Limit Theorem (CLT)*, which says that if a sufficiently large random sample is taken from some distribution, then even though this distribution is not itself approximately normal, the distribution of the sample *SUM* or *AVERAGE* will be approximately normal (more on this later)

Linear combinations of normals

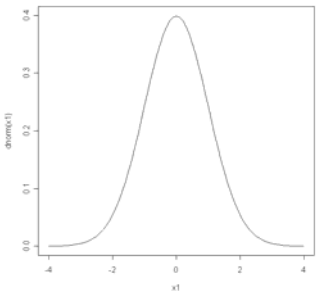
- An interesting and convenient fact: it turns out that a linear combination of normally distributed RVs is also normally distributed
- For example, consider $Z = aX + bY$, where
  - a and b are fixed numbers
  - $X \sim N(\mu, \sigma^2)$
  - $Y \sim N(\tau, v^2)$
- The distribution of Z is also normal, with mean = ?? and variance = ??

R: functions for normals

- Generate pseudo-random normals: `> rnorm(…)`
- Probability to the *left* of a value: `> pnorm(…)`
- Quantiles: `> qnorm(…)`
- (Height of the curve: `> dnorm(…)`)
- These 4 fundamental items can be computed for a number of common distributions (e.g. binomial, t, chi-square, etc.): `rbinom(), qt(), pchisq()...`

R: normal curve plot

```
> x1<-seq(-4,4,.1)
> plot(x1,dnorm(x1),
  type="l")
```



Example

- Suppose a RV X has a mean of 66 and SD of 9, and that X is approximately normally distributed
- Find the probability of obtaining a value between 57 and 75
- $P(57 < X < 75)$
- Find $P(X > 80)$

Finding normal quantiles

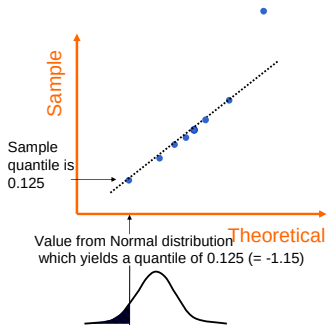
- The normal distribution can also be used to find *quantiles* when you know the probability
- In the previous problem, find the 75th percentile

Another example

- Among diabetics, the fasting blood glucose level may be assumed to be approximately normally distributed with mean 106 mg/100 ml and SD 8 mg/100 ml.
  - Find the chance of a level under 122 mg/100 ml
  - Find the chance of a level at least 122 mg/100 ml
  - About what percentage of diabetics have levels between 90 and 122 mg/100 ml?
  - Find the point x_0 that has the property that 25% of all diabetics have a fasting glucose level lower than x_0

QQ-Plot

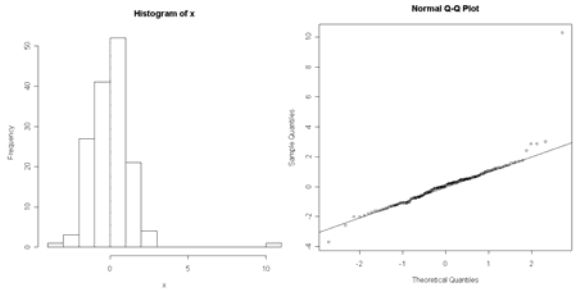
- Quantile-quantile plot
- Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)
- A method for looking for outliers when data are mostly normal



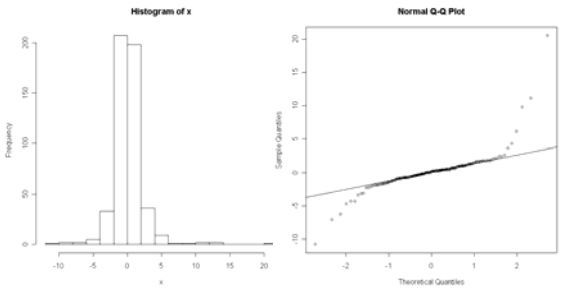
Typical deviations from straight line patterns

- Outliers
- Curvature at both ends (long or short tails)
- Convex/concave curvature (asymmetry)
- Horizontal segments, plateaus, gaps

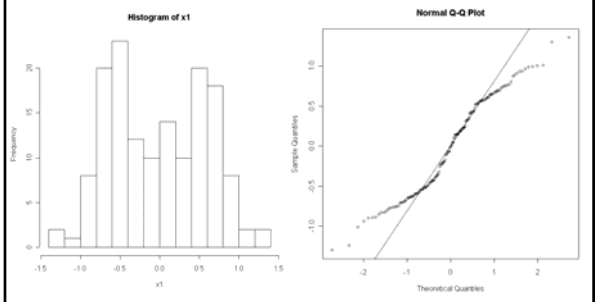
Outliers



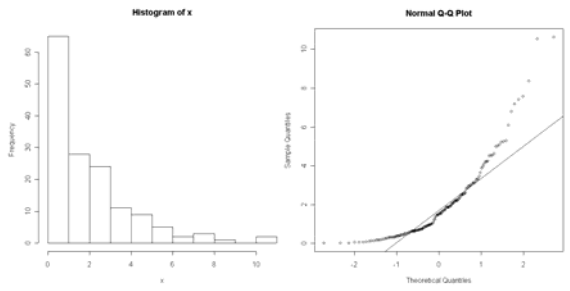
Long Tails



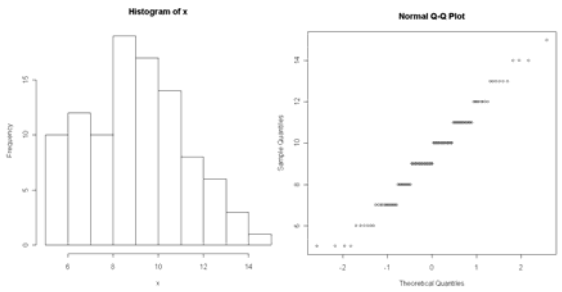
Short Tails



Asymmetry



Plateaus/Gaps



Sampling variability

- Say we sample from a population in order to estimate the population mean
- We would use the *sample mean* as our guess for the unknown value of the population mean
- Our sample mean is very unlikely to be exactly equal to the (unknown) population mean just due to *chance variation* in sampling
- If we estimate the mean *multiple times* from different samples, we will get a certain *distribution* of values

Central Limit Theorem (CLT)

- The *CLT* says that if we
  - repeat the sampling process many times
  - compute the sample mean (or proportion) each time
  - make a histogram of all the means (or proportions)
- then that histogram of sample means (or proportions) should look like the normal distribution
- Of course, in practice we only get one sample from the population
- The CLT provides the basis for making confidence intervals and hypothesis tests for means or proportions

Sampling variability of the sample mean

- Say the SD in the population for the variable is known to be some number σ
- If a sample of *n* individuals has been chosen 'at random' from the population, then the likely size of chance error of the sample mean (called the *standard error*) is

$$SE(mean) = \sigma / \sqrt{n}$$

- This the typical difference to be expected if the sampling is done twice independently and the averages are compared
- If σ is not known, you can substitute an *estimate*

Central Limit Theorem (CLT)

Sample 1	Mean 1
Sample 2	Mean 2
Sample 3	Mean 3
Sample 4	Mean 4
Sample 5	Mean 5
Sample 6	Mean 6
...	...

Normal distribution

Mean : true mean of the population
SD : $\sigma / \sqrt{n}$

Note: this is the SD of the sample mean, also called Standard Error; it is *not* the SD of the original population

Normal approximation to the binomial distribution

- One important application of the CLT is when the RV X ~ Bin(*n, p*) when *n* is large
- X is a sum of independent Bernoulli(*p*) RVs
- Then X is *exactly binomial*, but *approximately normal* with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$
- How large should *n* be? Large enough so that both *np* and *n(1-p)* are at least about 10
- Possible to modify the approximation if *np* or *n(1-p)* is between 5 and 10 (continuity correction)

Example

- A pair of dice is rolled approximately 180 times an hour at a craps table in Las Vegas
  a) Write an *exact expression* for the probability that 25 or more rolls have a sum of 7 during the first hour
  b) What is the *approximate probability* that 25 or more rolls have a sum of 7 during the first hour
  c) What is the *approximate probability* that between 700 and 750 rolls have a sum of 7 during 24 hours?

(BREAK)

Point estimation

- As opposed to (confidence) *interval* estimation
- Choose a *single value* (a 'point') to estimate an unknown parameter value
- We just looked at one method for doing this (ML)
- For concreteness, we will focus here on the problem of estimating the *population mean*
- Same *principles* apply for other parameters (but the details will be different)
- Generic parameter θ

Estimator properties

- What would be a *good way* to estimate the population mean based on a data set??
- Would like some *general principles* for comparing competing estimators
- We can look at *properties* like
  - bias
  - variance
  - mean square error (MSE)

Bias

- The *bias* of an estimator $\hat{\theta}$ for a parameter θ is defined as

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

i.e. the difference between the expected value of the sampling distribution of the estimator θ and the true value of the parameter θ

- An estimator is *unbiased* if the bias = 0

Bias: what does it mean?

- If an estimator is *unbiased*, it means:
  - Take a sample from the population, calculate the value of the estimator
  - do this many times
  - end up with a list of many sample estimates: make a histogram of these values
  - the average of this histogram is the same as the true (but unknown) population parameter value

Estimating the mean

- The sample mean is not the only possible estimator for the population mean μ
- It's not even the only unbiased estimator
- 'Lazy' estimator for the population mean: X_1 (just the first value, even though we have *n* of them)
- Another characteristic we can look at is the *variance* of the estimator

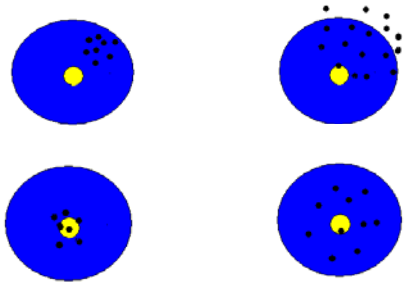
Variance

- The *variance* of an estimator $\hat{\theta}$ for a parameter θ is defined as

$$\text{Var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

- An estimator with lower variance is more *precise*
- Let's look at the variance of the sample mean and 'lazy' ...

Target practice



Mean square error

- Might want to consider an estimator with *some bias*
- Can compare estimators based on a *combination* of bias and variance called *mean square error* (*MSE*)

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

- It turns out that MSE can also be written

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$$

Sample surveys (review)

- *Surveys* are carried out with the aim of learning about characteristics (or *parameters*) of a *target population*, the group of interest
- The survey may select *all* population members (*census*) or only *a part* of the population (*sample*)
- Typically studies sample individuals (rather than obtain a census) because of time, cost, and other practical constraints

Introduction to CI Estimation

- Usually not very informative to give only a *point estimate* – a single value guess for the value of an unknown population parameter
- Better to present an estimate in the form of a *confidence interval* – a range of values for the parameter which seems likely given your sample
- To be concrete, consider CI for an unknown *population mean* (later for *population proportion*)
- CIs for other parameters have *different specifics*, but the *same ideas* and interpretations are behind them

CLT review

- The *CLT* says that if we
  - repeat the sampling process many times
  - compute the sample mean (or proportion) each time
  - make a histogram of all the means (or proportions)
- then that histogram of sample means (or proportions) should look like the normal dist. with
  - mean equal to the true population mean μ
  - SD equal to $\sigma/\sqrt{n}$ (σ is the SD for a single observation)
- The CLT provides the basis for making confidence intervals and hypothesis tests for means or proportions

Derivation of CI

- There is a 95% probability that the sample mean falls within 1.96 $\sigma/\sqrt{n}$ of the true mean μ:

  $P[\mu - 1.96\ \sigma/\sqrt{n} \le \bar{X} \le \mu + 1.96\ \sigma/\sqrt{n}] = .95$
- The event $\bar{X}$ being within 1.96 $\sigma/\sqrt{n}$ of μ is *the same event* as μ being within 1.96 $\sigma/\sqrt{n}$ of $\bar{X}$, so they have the same probability:

  $P[\bar{X} - 1.96\ \sigma/\sqrt{n} \le \mu \le \bar{X} + 1.96\ \sigma/\sqrt{n}] = .95$
- The random interval $(\bar{X} - 1.96\ \sigma/\sqrt{n},\ \bar{X} + 1.96\ \sigma/\sqrt{n})$ based on the observed sample mean is called a *95% confidence interval for* μ

CI for mean: mechanics

- When the CLT applies, a CI for μ looks like

  sample mean +/- $z^*\sigma/\sqrt{n}$,

  where z is a number from the *standard normal* chosen so the *confidence level* is a specified size (*e.g.* 95%, 90%, *etc.*)
- It's OK with me if you use 2 instead of 1.96...
- Let's find the z values for confidence levels: 68%, 90%, 99%, and any of your favorites ...

Example: mechanics

- Say we want to estimate μ = mean income of a particular population. A random sample of size n = 16 is taken; the sample mean is \$23,412, with an SD of \$2000.
- Estimate the population mean ...
- Make an approximate 95% CI for μ ...

Another example

- Say we want to estimate μ = mean exam score of a particular population. A random sample of size n = 25 is taken; the sample mean is 69.2, with an SD of 15.
- Estimate the population mean ...
- Make an approximate 90% CI for μ ...

Probability (but only a little)

- The *long run frequency interpretation* of *chance* or *probability* says that the chance of an event is the percentage (or proportion) of the time we expect the event to occur
- This is the most commonly used definition of probability, but is not the only one

CI for mean: interpretation

- **WRONG – WRONG – WRONG – WRONG**
- It is *tempting* - **BUT WRONG!!!** – to interpret a given 95% CI as saying that there is a 95% *chance* that the true parameter value is in the CI
- **WRONG – WRONG – WRONG – WRONG**
- Long-run frequency interpretation: there is *NO CHANCE* involved with the population mean μ
- μ is a *FIXED NUMBER*, we just don't know it
- Once the sample is drawn and the CI is fixed, then μ is either *IN* or *OUT* of that CI

So what does 95% mean?

- The 95% (for a 95% CI) is ***NOT*** the probability that a given CI contains the true μ
- The 95% part says something about the *sampling procedure*: if we did the whole procedure (get a sample of size *n* and make a 95% CI for the mean) over and over again, *about 95% of the intervals* made according to the (appropriate) mechanical rule would contain the true population mean μ
- Of course, in practice we don't obtain many samples of size *n*, we have just one – and we don't know if our interval is one of the 95% of 'good' ones or if it's in the 5% of 'bad' ones

Example

- The following data were obtained on a random sample of size 30 from the distribution of the percentage increase in blood alcohol content after a person drinks 4 beers:
  - sample mean = 41.2
  - sample SD = 2.1
- *Q:* Find a 80% CI for the (population) average percentage in blood alcohol content after drinking 4 beers.
- *A:* 41.2 +/- 1.28*(2.1/√30), or 40.7 to 41.7

Example, cont

- *Q:* Would a 95% CI be shorter or longer than the 80% CI we just made?
- *A:* (let's vote!)
- *Q:* If you hear a claim that the average increase is less than 35%, would you believe that claim?
- *A:* (let's discuss)

CI for population proportion

- For the population proportion, a 95% (say) CI is:

  sample proportion p +/- z*√[p(1-p)/n]

- Example: In a random sample of 36 graduate students at a particular large university, 8 have an undergraduate degree in mathematics. Find an approximate 95% CI for the proportion of graduate students at the university with undergraduate math degrees …
- Answer: assuming 36 is sufficiently 'large', the CI is .22 +/- 2*.07, or .08 to .36

A practice problem

- Acute myeloblastic leukemia is among the most deadly of cancers. Consider a RV X = the time in months that a patient survives after the initial diagnosis of the disease. Assume that X is normally distributed with a standard deviation of 3 months. Studies indicate that the mean μ = 13 months.
  - What is the chance that a randomly selected patient survives at least 16 months?
  - Suppose we have a random sample of 9 patients. Can we use the CLT to estimate the chance that the average survival of these 9 is at least 16 months? Why or why not, and if so compute this probability.
  - What is the 75th percentile for the survival time? For the average of 9 survival times?

Another practice problem

- To determine the effectiveness of a certain diet in reducing the amount of cholesterol in the blood stream, 100 people are put on the diet. After they have been on the diet for a sufficient length of time, their cholesterol count will be taken. The nutritionist running this experiment had decided to support the diet if at least 60% of the people have a lower cholesterol after going on the diet. What is the probability that the nutritionist supports the new diet if, in fact, it has no effect on the cholesterol level?

CI game

- Toss a die n = 4 times, make a 95% CI for the average value (σ = 1.7)
- Do this again, making a total of 5 CIs
- Now, toss 9 times, and make a 95% CI
- Again, make a total of 5 CIs
- Are you ready: try 25 times…
- Again, make a total of 5 of these CIs

- Yes, there is a point to all this! ☺