• Normal Approximation to the Binomial.
• Confidence Intervals: intuition and graphics.
• Confidence Intervals: formulas.

Normal Approximation to the Binomial

1. Sum of many independent 0/1 components with probabilities equal $p$ (with $n$ large enough such that $npq \geq 3$), then the binomial number of success in $n$ trials can be approximated by the Normal distribution with mean $\mu = np$ and standard deviation $\sqrt{np(1-p)}$.

2. For $n$ large, the sampling distribution of $\hat{p}$ can be approximated by a normal distribution with mean=$p$ and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

```
hist(rbinom(10000,20,0.5),xlim=c(0,20),
      probability=T,breaks=seq(0.5,20.5,1))
lines(seq(0,20,0.1),dnorm(seq(0,20,0.1),
       10,sqrt(5)))

#Non symmetric binomial
hist(rbinom(10000,20,0.3),xlim=c(0,20),
     probability=T,breaks=seq(-0.5,15.5,1))
lines(seq(0,20,0.1),dnorm(seq(0,20,0.1),
       6,sqrt(4.2)))
```

Continuity Correction:

$$P(a \leq X \leq b) \approx P\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

“Statisticians are the only people who insist on being wrong 5% of the time”
CONFIDENCE INTERVALS (S& W Chap 6)
Confidence interval for unknown \( \mu \) (with known \( \sigma \))
Interpretation of C.I.- repeated sampling and the confidence stack
What a confidence interval depends on: C, n and \( \sigma \)
Choice of sample size

Two Remarks to complement the last lecture on normal approximation and CLT:

1. Example: Consider incomes in town, where \( \mu = 39.97 \) and \( \sigma = 13.75 \): \( X_1 \) NOT normal.
   Sample, \( n=50 \), \( P(\bar{X}_{50} \geq 44) \)?
   \( \bar{X}_{50} \sim \mathcal{N}(39.97, \frac{13.75}{\sqrt{50}}) \)
   \( \bar{X}_{50} \) is approximately normally distributed with mean around 40 and sd 1.94,
   
   \[
P = P(\bar{X}_{50} \geq 44) = P(\frac{\bar{X}_{50} - 40}{1.94} > \frac{44 - 40}{1.94}) \approx P(Z > 2.06) = 2% \]

2. Remark. Adding independent variables brings the sum closer to being normal.
   Hence, if you start at the normal, you should stay there!
   If \( X \sim \mathcal{N}(\mu, \sigma^2) \) then \( \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \) exactly.

More generally, if \( X \) and \( Y \) are normal, independent, then \( aX+bY \) Normal
for any constants \( a, b \) (— a linear combination ). What are the mean & variance of \( aX+bY \) ?

Typical poll says “support for Bush is 52% with margin of error of 4%” This is an example of a confidence interval.
C.I.’s are one of the strangest animals in the statistical zoo, and one has to be careful with their interpretation. There has been quite a lot of philosophical debate about them, but nevertheless they remain a very useful tool for assessing the accuracy of estimates.

CONFIDENCE INTERVAL Estimate +/- Margin of Error: E +/- M

2 key components:

1) interval
   \( (E-M, E+M) \) (with estimate \( E \) at center)

2) confidence level 95%, 99% or other

\( C \) = Probability that the method yields an interval containing the true value (of the unknown parameter).

The confidence stack: Imagine drawing lots of samples – each generating a 95% C.I.
Some intervals do not overlap with the true value $\mu$, the randomness comes from the sample chosen NOT the mean which has a fixed unknown value.

**Examples:**

a) C.I. for population mean $\mu$, with **known** popn SD $\sigma$

b) C.I. for pop mean $\mu$, unknown $\sigma$.

c) C.I. for difference in two means, unknown $\sigma$.

**Preparation:** Book’s notation: $z_\alpha =$ location on standard normal curve with area $1 - 2\alpha$ under $(-z_\alpha, z_\alpha)$: quantiles
Conf. Interval for mean $\mu$, with known $\sigma$

Suppose a random variable $X$ has mean $\mu$ (unknown) and SD $\sigma$ (known), and that we have $n$ independent observations $x_1, x_2, \ldots, x_n$ of this r.v.

A level $C$, or $100(1 - 2\alpha)\%$ confidence interval for $\mu$ is $[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

The interval is “exact” if $X$ itself has a normal distribution approximately correct (by the CLT) for any $X$ if $n$ is large, usually we suppose $n > 20$.

**Standard error of the sample mean (and other sample statistics)**

If $\sigma$ known, then SD of sample mean, $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, when $\sigma$ is unknown, we use the estimated standard error of the mean:

$$s_\bar{x} = SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The sample mean is an example of a statistic $T$, (a quantity derived from a sample of data, such as $\bar{x}$). Other examples of statistics include the sample standard deviation $s$, sample coefficient of variation $CV$ sample skewness and kurtosis.

**Warning about names** for variability of random variables and statistics: Important to distinguish between the population value of the variability of a statistic, (which is generally unknown, since it depends on the whole population), and a sample estimate which is based on observed data from a probability sample. The latter is a random quantity (if we drew another sample, we would get a different estimate).

The term “standard error” is usually reserved for the SD of the sample mean The term “standard error of $T$” refers to the SD of a sample statistic $T$.

**Example** Confidence interval for the mean of IQs, for a population whose known variance is $\sigma^2 = 225 = 15^2$, Sample size $n=50$. $\bar{x} = 113.9$ observed mean. Special feature of IQs: normally distributed, and $\sigma = 15$ is known, so $C=95\%$, $z_{\bar{x}} = 1.96$ margin of error $M = 1.96 \times 15 / \sqrt{50} = 1.96 \times 2.12 = 4.2$

95\% CI is $[113.9 - 4.2, 113.9 + 4.2] = [109.7, 118.1]$

A level $C$, or $100(1 - \alpha)\%$ confidence interval for $\mu$ is

$$[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

But to return to reality, we don’t know $\sigma$. Thus we must estimate the standard deviation of $\bar{X}$ with:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

But $s$ is just a function of our $X_i$’s and thus is a random variable too – it has a sampling distribution too.

Before we could say if we knew $\sigma$

$$P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

which after algebra gave the confidence interval.

[Remember for any $s$, $z_s$ is defined as where $1 - 2s$ of the area falls in $(-z_s, z_s)$. So $z_s = qnorm(1 - s) = -qnorm(s) = 1 - s$ quantile. i.e. $z_s$ is the positive side.]

Now we want a similar setup, so that:

$$P(?? < \frac{\bar{X} - \mu}{SE_{\bar{X}}} < ??) = \alpha$$

We need know the probability distribution of $T = \frac{\bar{X} - \mu}{SE_{\bar{X}}}$. $T$ has the Student’s t-distribution with $n - 1$ degrees of freedom. We write this as $T \sim t_{n-1}$. The degrees of freedom=$\nu$ is the only parameter of this distribution.