18. Sampling distribution models

- Sampling variability, sampling error.
- Central Limit Theorem
- Sampling distribution model for a statistic:
  - for a proportion
  - for a mean

Where are we?

- Ch. 2–9: Data set and its distribution, statistics, Normal model.
- Ch. 11–13: Collecting data, random samples, randomized experiments.
- Ch. 14–16: Probability, random variables, probability distribution models.
- Ch. 18–28: Statistical inference: what does sample data tell us about the underlying population. Inferences about parameters (proportions, means, etc.) in a model for the population distribution.
- First step: concept of a sampling distribution model for a statistic.

- Drop a thumb tack on the floor. What is the probability p that the thumb tack lands with the point up?

- Toss the thumb tack n times. Let $\hat{p}$ be the sample proportion. How close will $\hat{p}$ be to p?

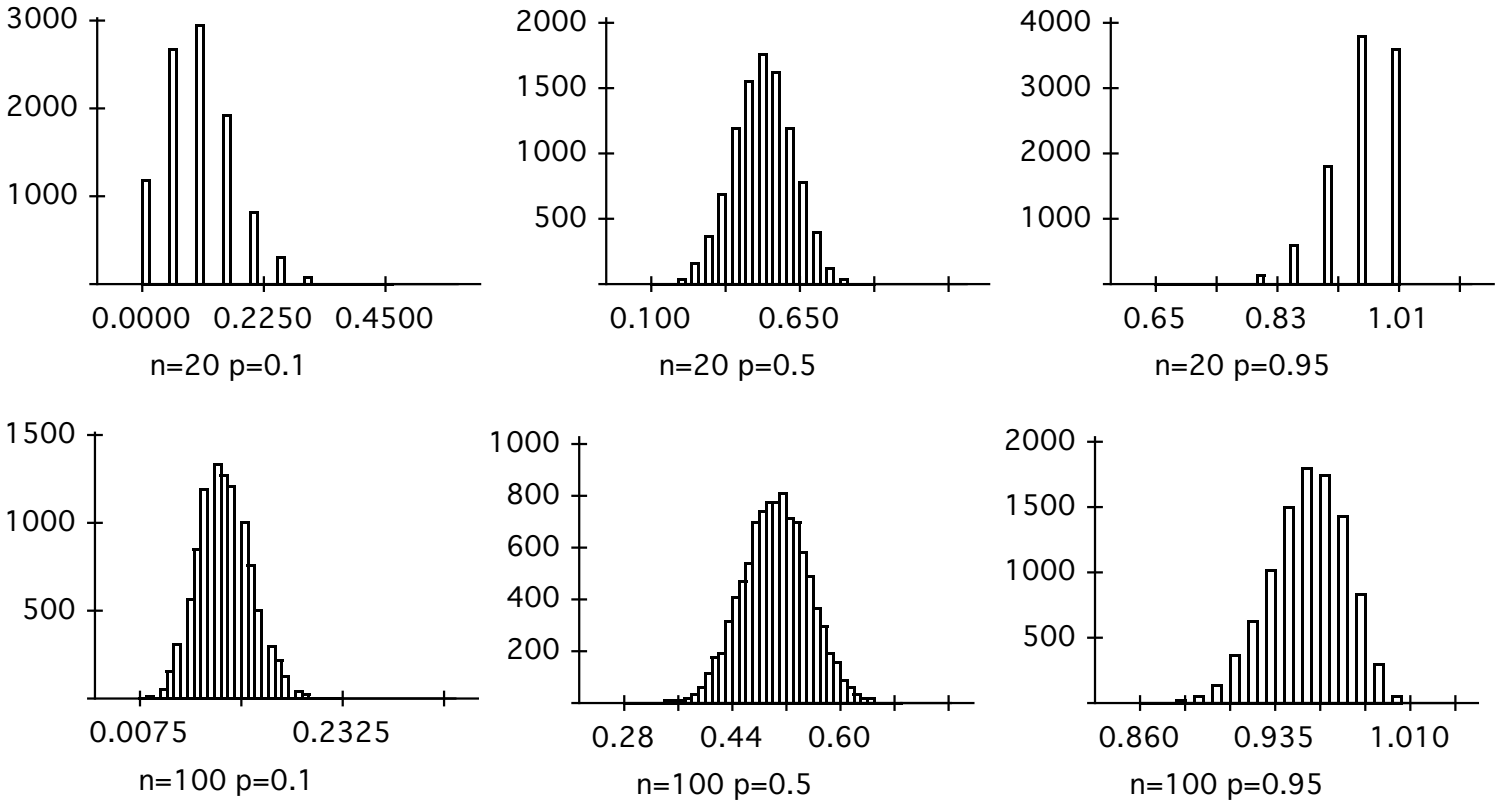
- Statistical model: n independent trials, n is fixed, constant probability p of "success", $\hat{p} =$ proportion of successes. Notation: p is a parameter, $\hat{p}$ is a statistic. (Some authors use π and p.) Put $q = 1 - p$.

- Key idea: We only take one sample, but we can imagine taking many samples of size n and getting many values of $\hat{p}$. The resulting distribution of these imaginary $\hat{p}$ values is called the sampling distribution.

- What can we say about the location, spread, and shape of the sampling distribution?

- If we knew the value of p, we could simulate values of $\hat{p}$.

Histograms based on 10000 observations of sample proportions.



- Extensive simulations would suggest that

$$E(\hat{p}) = p \quad \text{and} \quad SD(\hat{p}) = \sqrt{\frac{pq}{n}}\,.$$

  These formulae can be derived from rules for means and variances.

- The shape of the sampling distribution depends on n and p.

- When n is small, the shape depends on p:
  - Symmetric when $p = 0.5$.
  - Skewed to the right when p is close to zero.
  - Skewed to the left when p is close to one.

- When n is sufficiently large, the shape is approximately Normal.

- Central Limit Theorem for proportions:
  - Sampling distribution of $\hat{p}$ approaches Normal as n grows.
  - Justifies Normal sampling distribution model $N(p, \sqrt{\frac{pq}{n}})$.

- In practice, we need to verify three conditions:
  - Randomization condition: we have a simple random sample or a randomized experiment.
  - 10% condition: When sampling without replacement from a population, the sample size n should be at most 10% of the population size.
  - Success/failure condition: np and nq should both be at least 10. Note that np is the expected number of successes and nq is the expected number of failures.

- Thumbtacks: Someone claims that $p = 0.60$. Which of the following results seems consistent with this claim? State assumptions.

  1. Obtain $\hat{p} = 0.50$ with $n = 10$ tosses.
  2. Obtain $\hat{p} = 0.50$ with $n = 30$ tosses.
  3. Obtain $\hat{p} = 0.50$ with $n = 100$ tosses.

- Is the Normal sampling distribution model for $\hat{p}$ appropriate in the following situations.

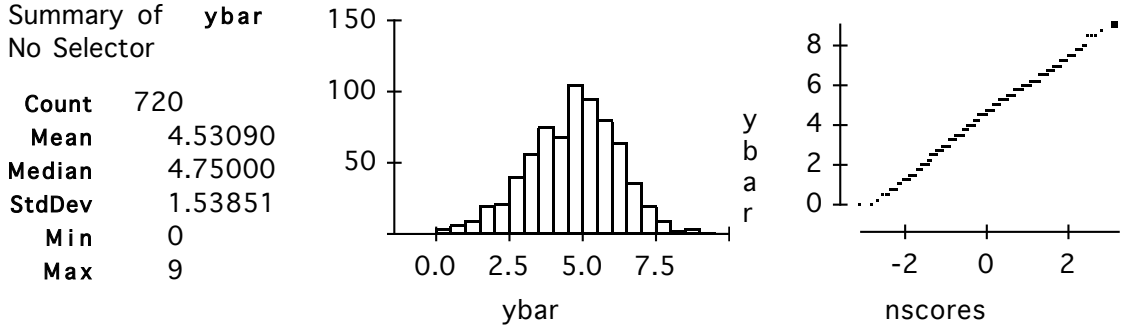
  1. The mayor of Edmonton takes a simple random sample of size 1000 from residents of the city to estimate support for a proposed policy. It is expected that support will be around 60%.
  2. The mayor of Jasper (pop $\approx$ 4200) does the same.
  3. A candidate for the Outer Fringe Party takes a simple random sample of 1000 Alberta voters to estimate support. The actual proportion supporting the candidate is well below 1%.
  4. An internet news site invites readers to indicate their preference in the federal election and receives 13,211 responses with 5284 supporting the Conservatives.

- *The Book of Risks* (Laudan, 1994) states that "each occasion of unprotected sex between a fertile, healthy couple of childbearing age poses a 5% to 10% risk of pregnancy". Suppose the "risk" for a particular population is 8%. If 250 couples from this population have unprotected sex on February 14, what is the probability that at least 20 pregnancies result? Assumptions?

So much for proportions, what about means?

- Let $\bar{Y}$ be the average of the last four digits from a randomly sampled telephone number. Simple model: the four digits are sampled randomly and independently from $\{0, 1, 2, \ldots, 9\}$. The model for each randomly sampled digit has $\mu = 4.5$, $\sigma^2 = 8.25$, and $\sigma \approx 2.872$.

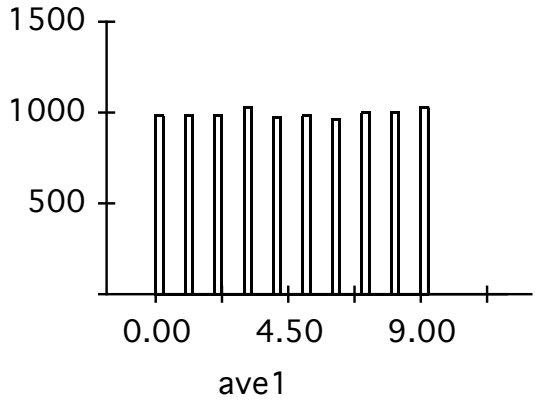
- What can we say about the center, spread, and shape of the sampling distribution of $\bar{Y}$? Investigate by (a) collecting data from class and (b) simulating data from assumed model.

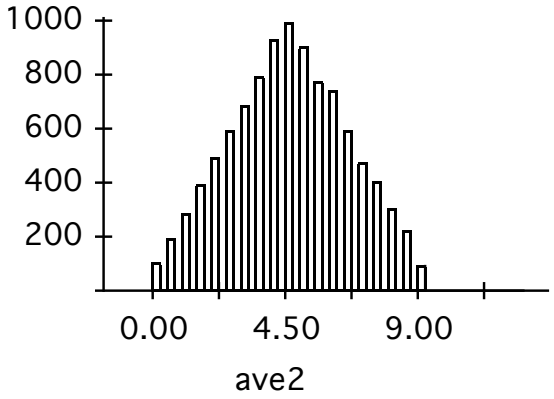


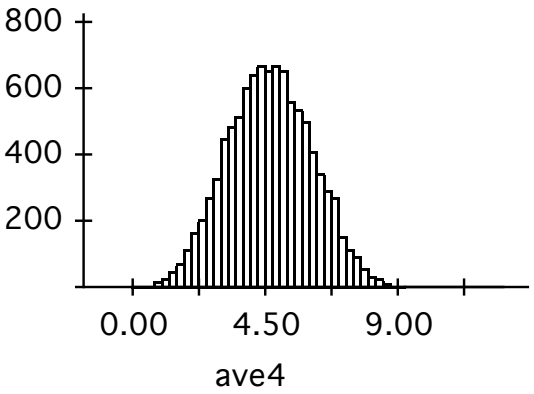
Generate 10000 replicates of 10 variables uniformly distributed on the digits {0, 1, 2, ..., 9}.
Define ave1, ave2, ave4, and ave10 as the average of the first 1, 2, 4, and 10 variables.

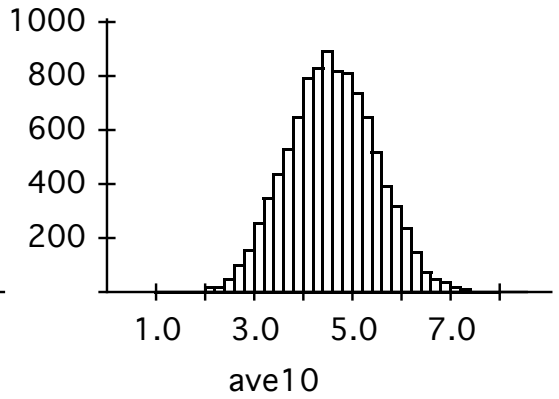
Summaries
No Selector

Variable	Mean	StdDev
ave1	4.51320	2.88043
ave2	4.51840	2.03198
ave4	4.52733	1.42642
ave10	4.52602	0.903705









- Simple random sample of size n from a much larger population
  Population model parameters μ and σ
  $\bar{Y}$ = sample mean

- The mean and standard deviation of the sampling distribution are

$$E(\bar{Y}) = \mu \quad \text{and} \quad SD(\bar{Y}) = \frac{\sigma}{\sqrt{n}} \, .$$

- Central Limit Theorem (Laplace, 1810):
  Sampling distribution of $\bar{Y}$ approaches Normal as n grows.

- Conditions underlying Normal sampling distribution model for $\bar{Y}$:
  - Randomization condition.
  - 10% condition.
  - Sufficiently large n condition. Larger n for skewed or heavy-tailed populations. But $n = 1$ sufficient if population model is Normal.

- CLT assumes (almost) nothing about the population model.
- Special case of CLT for proportions: observations in $\{0, 1\}$.
- Law of Diminishing Returns: Suppose the population model has $\sigma = 10$.
  - What is the standard deviation of the sampling distribution model for $\bar{Y}$ for each of the following sample sizes n: 1, 4, 9, 16, 25, 100.
  - How large a sample size is needed so that the sample mean has standard deviation at most 0.1?

- The distribution of lengths of trout fry in a pond at a fish hatchery is approximated by a Normal model with mean 8.6 cm and standard deviation 2.0 cm. A dozen fry will be netted and their lengths measured.

  a) What is the probability that the average length of the 12 fry will be less than 7.6 cm?

  b) Do you think the netted fry would represent a random sample? If not, how would this affect your answer in (a)?

- The mean of all credit card balances for I.O.U. Credit Corporation is \$300, and the standard deviation is \$180. Let $\bar{Y}$ be the average balance for a random sample of 81 credit cards.

  a) Find the mean and standard deviation of $\bar{Y}$.

  b) What can you say about the distribution of $\bar{Y}$.

  c) Find $P(280 \leq \bar{Y} \leq 320)$.

  d) Can you find $P(280 \leq Y \leq 320)$, where Y is the balance for a single randomly selected credit card? How would this probability compare with that in part (c).

- You are buying n items at a grocery store. To get a rough idea of the total cost, you round off the price of each item to the nearest dollar, then add. How close should your answer be to the correct amount?
  - Let $Y_1, Y_2, \ldots, Y_n$ denote the individual rounding errors. E.g., if price $= 2.35$ then $Y = 0.35$, if price $= 5.87$ then $Y = -0.13$.
  - Total error $= Y_1 + \ldots + Y_n = n\bar{Y}$.
  - Simple model: $\{Y_1, \ldots, Y_n\}$ is a random sample from a population with uniform distribution on interval $(-0.5, 0.5)$, parameters $\mu = 0$, $\sigma = 1/\sqrt{12} = 0.289$.
  - If you purchase 20 items, what is the probability that your total cost estimate is within $\$2.00$ of the correct amount?

19. Confidence intervals for proportions

- Margin of error = (critical value) $\times$ (standard error)
- Confidence level related to sampling distribution model.
- Three conditions needed for model.
- Certainty versus precision.
- Choosing your sample size.
- One-proportion z-interval compared with Plus-four interval.

- Let p be the long run probability that a thumbtack will land with the point up. You toss the thumbtack $n = 50$ times and it lands with point up 27 times. How close is the unknown p to the the estimate $\hat{p} = 27/50$?

- Suppose the Normal sampling distribution model holds; i.e., $\hat{p}$ is Normal with $E(\hat{p}) = p$ and $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$. Is that reasonable? Conditions? By the 68-95-99.7 rule, $\hat{p}$ will be in the interval $p \pm 2\,SD(\hat{p})$ for about 95% of samples. And that's the same as p being in the interval $\hat{p} \pm 2SD(\hat{p})$.

- But we have a problem: $SD(\hat{p})$ is unknown. Solution: Replace SD by a statistic called the *standard error*:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

  Under the Normal sampling distribution model, p will be in the interval $\hat{p} \pm 2\,SE(\hat{p})$ for about 95% of samples.

- For our particular sample, we have $\hat{p} = 27/50 = 0.54$ and $SE = 0.070$.

- Can we conclude p has a 95% probability of being in the interval 0.54 ± 0.140? I.e., can we write $P(0.40 \leq p \leq 0.68) = 0.95$?

- No, p is a parameter, not a random variable. The parameter is either in the interval or not.

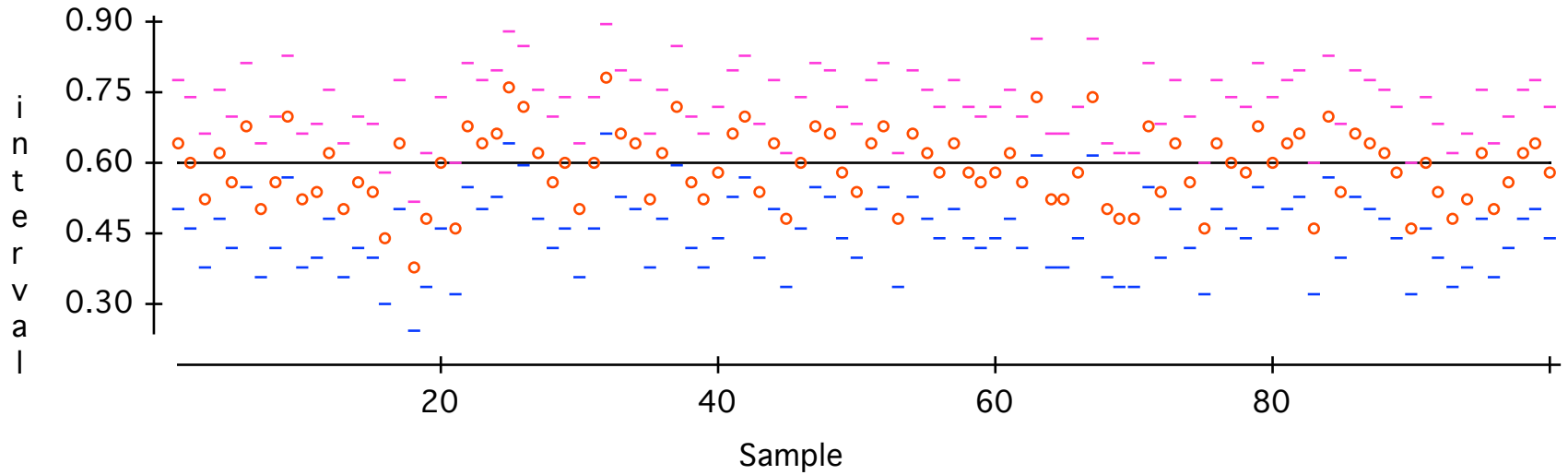
- We *can* say that the method used to obtain the data and construct the interval produces an interval containing p for about 95% of samples.

- More briefly, we have 95% *confidence* that p is between 0.40 and 0.68.

- In general, *confidence* refers to the probability that an interval covers a parameter, where the probability is calculated and interpreted using a sampling distribution model.

- Terminology: confidence level $= 95\%$, margin of error $= ME = 2\,SE$, confidence interval is $estimate \pm ME$.

- 100 confidence intervals for p simulated with $n = 50$ and $p = 0.60$.



- There are many ways to misinterpret a confidence interval.

- In a national poll, 1000 Canadians were asked which party they planned to vote for in the federal election and 380 indicated support for the Conservatives. Confidence interval?

  a) Think: Verify 3 conditions for the Normal sampling distribution model.
  b) Show: Calculate $\hat{p}$, SE, ME, 95% confidence interval.
  c) Tell: Interpret result.

- In the same poll, 67% of Albertans polled indicated support for the Conservatives. Confidence interval?

- Margin of error: certainty versus precision.
  - Make interval smaller by decreasing the confidence level.
  - Make the confidence level larger by increasing the margin of error.

- Given a confidence level, define the *critical value* z^* so that $P(-z^* < Z < z^*) =$ confidence level. Here Z is standard Normal.

Confidence level	Critical value
80%	1.28
90%	1.645
95%	1.96 $\approx$ 2
99%	2.58

Margin of error = (critical value) $\times$ (standard error)

$$ME = z^* \times SE(\hat{p})$$

- The interval $\hat{p} \pm ME$ is called the *one-proportion z-interval*.

- In a pilot study with small n, we might be satisfied with 80%.
  In a large study, we might want 99%.

- The confidence level should be chosen before looking at the data.

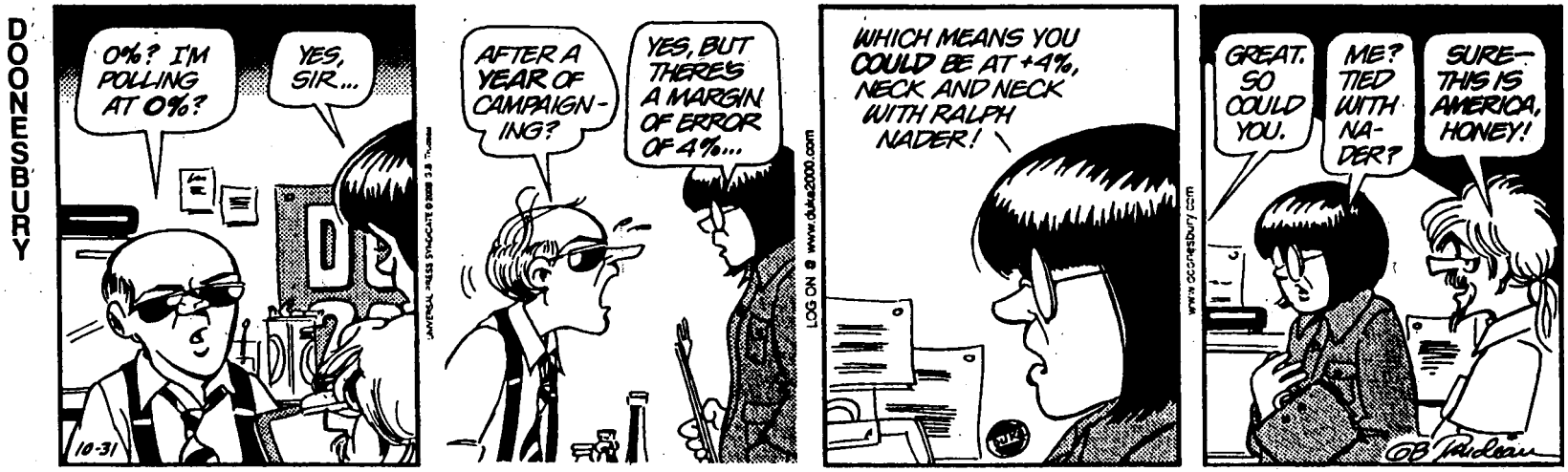
- The confidence level is usually one of the values in the table.
  An unusual value might raise suspicions.

- Compare intervals in the national poll example.

- When news media report poll results, they usually use 95% confidence level and a "worst case" value for the standard error; i.e., they report

$$ME = (1.96)\sqrt{\frac{(0.5)(0.5)}{n}} = \frac{1.96}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

as a measure of precision for all proportions reported in the poll. This can lead to silly conclusions when $\hat{p}$ is near 0 or 1.



- If the media report that its poll results are "accurate to within three percentage points 19 times out of 20", what was the sample size?

- If 3000 persons were included in the poll, what margin of error would the media report?

- The margin of error reported in the media refers to proportions of the entire sample, not subgroups (e.g., regions).

- Choosing your sample size: Decide on ME and confidence level, then calculate z^*. Recall:

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Solve for n:

$$n \approx \hat{p}\hat{q}\left(\frac{z^*}{ME}\right)^2$$

Substitute a value for $\hat{p}$ (and $\hat{q} = 1 - \hat{p}$), then round up. This value for $\hat{p}$ is just used to determine n.

- If you expect p to be somewhere in the middle, use the "worst case" value of $\hat{p} = 0.5$ to ensure ME is sufficiently small for all values of p.

- If you believe p is less than some small upper bound, then you can set $\hat{p}$ equal to the upper bound. Here a relatively small ME is usually desired.

- You are planning an opinion poll and want the results to be accurate to within two percentage points 19 times out of 20. Sample size?

- A bank wants to test market a new credit card. It will mail out information about the card to a random sample of persons, inviting them to return an application form if they are interested. The bank wants to estimate the proportion of its target population that would be sufficiently interested to return an application. The bank is fairly sure that at most two percent would be interested. It wants ME no greater than one quarter of a percentage point with 99% confidence. Sample size?

- You select a random sample of size 100 and observe 12 successes. Is the usual confidence interval appropriate?

- Agresti-Coull or "Plus-four" interval. Add four phony data points, two successes and two failures, then proceed as before.

$$\tilde{p} = \frac{\#\text{ successes} + 2}{n+4} \quad \text{and} \quad \tilde{q} = 1 - \tilde{p} = \frac{\#\text{ failures} + 2}{n+4}$$

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}\tilde{q}}{n+4}}$$

  The interval is centered around $\tilde{p}$, not $\hat{p}$. This is reasonable because the sampling distribution of $\hat{p}$ is perfectly symmetric only when $p = 0.5$.

- The Plus-four interval is strongly recommended when the Success/Failure condition is questionable. It can also be used as a general replacement for the one-proportion z-interval, but not on MathXL assignments.

- Compare the Plus-four interval with the one-proportion z-interval assuming $n = 50$, $\hat{p} = .12$, and confidence $= 95\%$.

- Mad Cow: In 2002, 3.45 million cattle were slaughtered in Canada. Among these, 3377 were tested for BSE. No infected cattle were found. Suppose the tested cattle were a simple random sample from a sub-population believed to be at higher risk. Let p denote the proportion of infected cattle among the sub-population. Confidence interval?

- Don't misstate what the interval means.
  - Don't suggest that the parameter varies.
  - Don't claim that other samples will agree with yours.
  - Don't be certain about the parameter.
  - Don't forget, its about the parameter.
  - Don't claim too much much (e.g., about scope of study).

- Remember:
  - Take responsibility for uncertainty.
  - Treat the whole interval equally.
  - Reduce confidence level if margin of error is too large to be useful. This should be done during the planning stage.
  - Think about assumptions needed for the sample distribution model, especially independence.
  - Watch out for biased samples.

20. Testing hypotheses about proportions

- Null hypothesis H_0, a statement about model parameters.
- Test: decide whether H_0 is consistent with observed data.
- Alternative hypothesis H_A, one-sided or two-sided.
- Test statistic and sampling distribution model given H_0.
- P-value: a measure of the evidence against H_0 in the direction of H_A.
- Reject H_0 or fail to reject H_0, never accept H_0.
- One-proportion z-test.

- Satisfied customers? A large firm claims that 90% of its customers are satisfied with the firm's services. You think the claim may be an exaggeration and poll a simple random sample of the firm's customers. Of the 150 responses, 129 indicate satisfaction. Are you justified in rejecting the claim?

- Let p be the actual proportion of satisfied customers, a parameter.

- Null hypothesis $H_0 : p = 0.9$
  - Null means nothing or naught, so H_0 is pronounced H-naught.
  - In general, a null hypothesis makes a statement about some property of a model, often in the form: $parameter = value$.
  - H_0 usually represents "no change from traditional value", "no effect", "no difference", or "no relationship".

- Test: Decide whether H_0 is consistent with observed data.

- Decision depends on the alternative hypothesis H_A
  - H_A is a statement about the model that seems plausible (a priori) if H_0 is not true.
  - Hypotheses are formulated based on the context of the problem, not on the data used to carry out a test.
  - Here we have a one-sided alternative $H_A : p < 0.9$. Why?

- Criteria for decision? Consider what values of $\hat{p}$ are likely given H_0 and what values are likely given H_A. If the observed value is unlikely under H_0 and much more likely under H_A, then reject H_0.

- So you will reject H_0 if $\hat{p}$ is substantially smaller than 0.9. What does "substantially" mean?

- If H_0 is true, what can we say about the sampling distribution of $\hat{p}$? Verify the three conditions: randomization, 10%, success/failure. If these hold, then the sampling distribution of $\hat{p}$ is Normal with

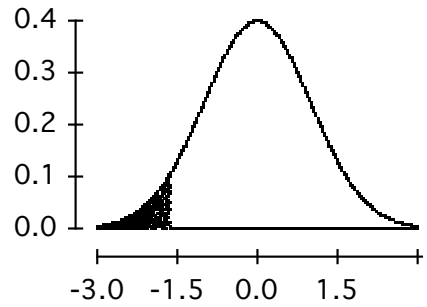
$$E(\hat{p}) = 0.9 \quad \text{and} \quad SD(\hat{p}) = \sqrt{\frac{(.9)(.1)}{150}} = 0.0245 \ .$$

- Our test statistic is obtained by standardizing $\hat{p}$ using the mean and standard deviation under H_0:

$$z = \frac{\hat{p} - E(\hat{p})}{SD(\hat{p})} = \frac{\hat{p} - 0.9}{0.0245} \ .$$

  The sampling distribution model for z is standard Normal.
  Note: we use the standard deviation here, not the standard error.

- We have $\hat{p} = 129/150 = 0.86$ so $z = -1.63$. Use the standard Normal model to calculate the P-value$= P(Z < -1.63) = 0.0516$.



- A P-value is a probability calculated using the sampling distribution model for our test statistic assuming H_0 is true. It is the probability of obtaining a value at least as unusual as the test statistic value actually observed. In this context, values are more "unusual" if they would be more likely to occur under H_A.

- A P-value is a measure of evidence against H_0 in the direction of H_A. Smaller P-values represent stronger evidence against H_0.

- Given P-value $= 0.0516$, should we reject H_0 or not?
  - What action would that entail?
  - What are the consequences if reject H_0 and $p \geq 0.9$?
  - What are the consequences if we fail to reject and $p < 0.9$?
  - How strong were your prior beliefs about the hypotheses?
  - Is a confidence interval helpful here?
  - Is a decision really required?

- What decision would you make for the following outcomes ($n = 150$):

Satisfied	123	125	127	128	129	130	135	140	145
$\hat{p}$	0.820	0.833	0.847	0.853	0.860	0.867	0.900	0.933	0.967
z	–3.27	–2.72	–2.18	–1.91	–1.63	–1.36	0.000	1.36	2.72
P-value	0.001	0.003	0.015	0.028	0.051	0.087	0.500	0.913	0.997

- If you had observed $\hat{p} = 0.900$, would you accept H_0?
  If you had observed $\hat{p} = 0.967$, would you reject H_0?

- Analogy: Criminal trial under British common law.
  - H_0: Defendant is innocent.
  - H_A: Defendant is guilty.
  - Data: Evidence deemed admissible and presented in court.
  - Small P-value: Proof of guilt beyond a reasonable doubt.
  - Decision: Guilty or Not guilty.

- Science
  - Scientific theories are falsifiable.
  - No theory can be proven true.
  - Science is organized common sense where many a beautiful theory was killed by an ugly fact. Thomas Henry Huxley (Darwin's "bulldog")
  - Is "intelligent design" a scientific theory?

- Four steps for hypothesis testing.

  1. Hypotheses: context, parameters, H_0, H_A.
  2. Model: name of test, formula for test statistic, sampling distribution model under H_0 (remember to verify conditions), formula or picture for calculating P-value.
  3. Mechanics: calculate test statistic and P-value.
  4. Conclusion: report your personal decision based on P-value with an interpretation in context.

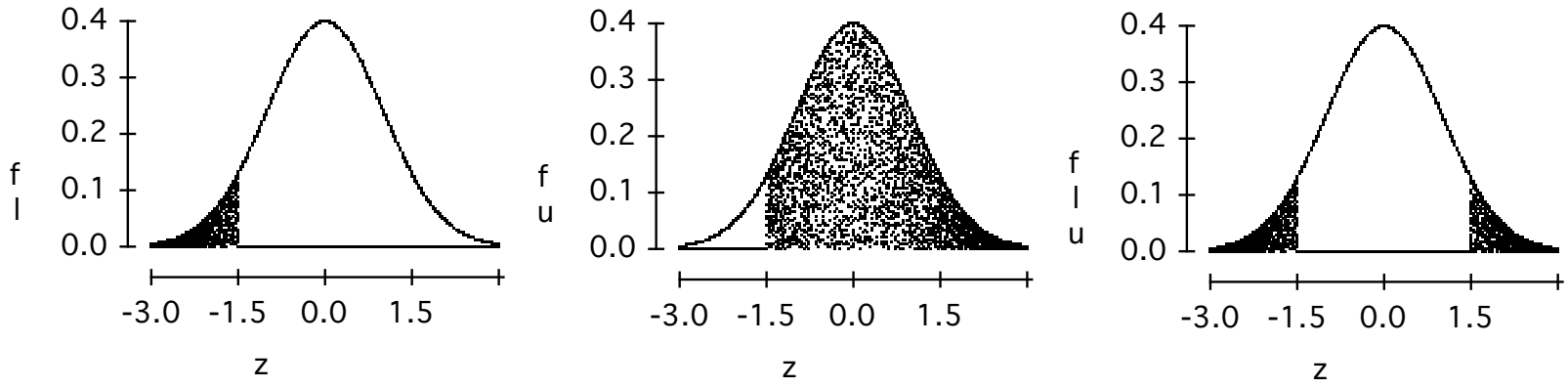
- Note that the "Think" steps 1 and 2 do not involve the sample data.

- Has support for the Alberta government changed since the last provincial election. The Conservatives received 52.7% of the vote. Suppose a poll with $n = 750$ shows support currently at 50%.

  $H_0 : p = 0.527$, $\hat{p} = 0.500$, $SD(\hat{p}) = 0.0182$, $z = -1.48$.

  Which alternative hypothesis and corresponding P-value is appropriate?

  $H_A : p < 0.527$ $\qquad$ $H_A : p > 0.527$ $\qquad$ $H_A : p \neq 0.527$



- A gambler has been studying a roulette wheel. If the wheel is out of balance, he can improve his odds of winning. Of the 38 numbers on an American roulette wheel, 18 are red, 18 are black, and 2 are green. If the wheel is balanced, the probability of the ball landing on red is $18/38 \approx 0.4737$. The gambler observes 200 spins of the wheel and finds that the ball lands on red 93 times. Is there evidence that the wheel is out of balance?

- What can go wrong?
  - Don't base H_0 on what you see in the data.
  - Don't base H_A on the data, either.
  - Don't make your null hypothesis what you want to show to be true.
  - Don't forget to check conditions.
  - If you fail to reject the null hypothesis, don't think that a bigger sample would be more likely to lead to rejection.

21. More about tests

- Hypotheses, P-values, decisions, context.
- Significance level α, critical value z^*.
- Connections between tests and confidence intervals.
- Type I and Type II errors, power of a test, trade-off.
- Statistical versus practical significance, effect size.

- Hypotheses, P-values, decisions, and context.
  - A women says she can tell by taste alone whether milk is poured into the cup before or after the tea.
  - Has support for a carbon tax changed over the past year?
  - Is there a home team advantage?
  - In North American football, does "freezing the kicker" work?

- More about P-values.
  - A P-value is a conditional probability; i.e., the probability of getting a result at least as unusual as the observed test statistic, given that H_0 is true.
  - A P-value is *not* the probability that the null hypothesis is true, given the data. This is a very common error.
  - $P(\boldsymbol{A} \mid \boldsymbol{B})$ is usually not the same as $P(\boldsymbol{B} \mid \boldsymbol{A})$.
  - You can think of a P-value as a probability concerning a random value of the test statistic obtained from a second hypothetical sample drawn independently from the observed sample.

- What to do with a high P-value.
  - A small P-value (i.e., close to 0) is interpreted as evidence against H_0 in the direction of H_A.
  - A high P-value (ı.e., close to 1) should *not* be interpreted as evidence supporting H_0. See "satisfied customer" example from Ch 20.
  - A high P-value for a one-sided test might suggest a reformulation; e.g., test $H_0 : p \geq 0.9$ versus $H_A : p < 0.9$. This reformulation leads to the same practical conclusion as $H_0 : p = 0.9$ versus $H_A : p < 0.9$. For simplicity, our null hypotheses have the form $H_0 : p = p_0$.
  - What if we are testing $H_0 : p = 0.5$ versus $H_A : p \neq 0.5$ and we obtain $\hat{p} = 0.5$. What is the P-value? Can we accept H_0 now?

- Alpha levels (aka significance levels): Sometimes our test involves making a firm decision or action; e.g., reject a batch of widgets if too many are faulty. In this situation, it is desirable to have a clear strategy:
  - Choose a value α before looking at the data.
  - Reject H_0 if P-value $< \alpha$, otherwise fail to reject.

- The decision can also be carried out by comparing the test statistic z with a critical value z^*. Assume $H_0 : p = p_0$. Let $Z \sim N(0, 1)$.

H_A	Reject H_0 if	z^* determined by
$p < p_0$	$z < z^*$	$P(Z < z^*) = \alpha$
$p > p_0$	$z > z^*$	$P(Z > z^*) = \alpha$
$p \neq p_0$	$\lvert z \rvert > z^*$	$P(Z > z^*) = \alpha/2$

- Some traditionally used values for $\alpha : 0.001,\ 0.01,\ 0.05,\ 0.10$.

- Two-sided tests and confidence intervals. Given the data:
  - a test tells you whether a particular parameter value seems plausible;
  - a confidence interval tells you which parameter values seem plausible.

- Consider testing $H_0 : p = p_0$ versus $H_A : p \neq p_0$ at level α. The $100(1-\alpha)\%$ confidence interval for p is (almost) the same as the interval of values p_0 for which H_0 would not be rejected.

- E.g., $n = 500$, $\hat{p} = 0.440$, 95% confidence interval is $(0.396, 0.484)$.

p_0	0.380	0.390	0.400	0.410	0.470	0.480	0.490	0.500
P-value	0.006	0.022	0.068	0.173	0.179	0.073	0.025	0.007

- This relationship between tests and confidence intervals is approximate here, not exact, because tests use $SD(\hat{p})$ while intervals use $SE(\hat{p})$.

- Who'll stop the rain. (Ex 20.32). A study of the effects of acid rain on trees in the Hopkins Forest shows that 25 of 100 trees sampled exhibited some sort of damage from acid rain. This rate is higher than the 15% quoted in a recent *Environmetrics* article on the proportion of damaged trees in the Northeast. Is there convincing evidence that the trees in the Hopkins Forest are more susceptible than trees from the rest of the region. Address this question with both a test and a confidence interval.

- Consider a claim that certain individuals have psychic abilities allowing them to remotely gather information.
  - How would you design a study to test this claim?
  - Number of trials?
  - Criteria for decision?
  - Potential errors and their implications?

- Errors, alpha level, and power (see demo).

	H_0 True	H_A True
We reject H_0	Type I Error	OK
We fail to reject H_0	OK	Type II Error

$$\begin{aligned} \alpha &= \text{probability of Type I Error} \\ \beta &= \text{probability of Type II Error} \\ 1-\beta &= \text{power of the test} \end{aligned}$$

  - The significance level α is known, chosen before observing data.
  - The power increases with the *effect size*; i.e., the absolute difference between the actual and null values of the parameter.
  - Trade-off between α and power, similar to trade-off between confidence level and margin of error.
  - To increase power while keeping α fixed, we need to increase n.

- Analogy: Law and Order
  - In a criminal trial, the defendant is presumed innocent until guilt is proven beyond a reasonable doubt.
  - In a civil trial, the object is to determine a level of responsibility.

- Analogy: Diagnostic test for a medical condition
  - A test result is "positive" if the condition appears to be present. Otherwise the test result is "negative".
  - Sensitivity = proportion of positive test results among subjects who have the condition.
  - Specificity = proportion of negative test results among subjects who do not have the condition.

- When planning a study and choosing a significance level α:

- Take into account the implications (costs) of the two types of error. E.g., better to have a false positive than a false negative.

- Choose n large enough so that the power is sufficiently large for an effect size of practical importance. Pilot studies are often used for this purpose.

- Given n, keep the power in mind when choosing α.

- The choice $\alpha = 0.05$ is widely used, but apparently with little justification. Fisher suggested $\alpha = 0.05$ might be reasonable in some situations. It appears he had in mind small studies, where a much smaller value of α would produce a test with little power.

- Two kinds of significance: statistical (alpha level) and practical.
- Statistical significance: The observed effect seems to be too large to have occurred by chance if the null hypothesis were true.
- Practical significance: The effect size is large enough to be important. This depends on the unknown true value of the parameter and on the context. E.g., toxicity, election result.
- When n is large, an observed effect may have statistical significance but lack practical significance.
- When n is small, the true effect size may have practical significance even though the observed effect lacks statistical significance.
- It is often helpful to report both a P-value and a confidence interval.

22. Comparing two proportions

- Two independent samples or randomized experiment.
- Confidence interval for $p_1 - p_2$. Two-proportion z interval.
- Variance of $\hat{p}_1 - \hat{p}_2$ equals sum of variances.
- $H_0 : p_1 - p_2 = 0$. Two-proportion z-test.
- The test uses a pooled estimate of proportion given H_0.
  The interval uses individual estimates.

Example. APGAR score and gestational age at birth.

- APGAR scores provide measures of health for newborn infants. Low scores indicate problems. A 5-minute APGAR less than 7 provides a relatively stable indicator of poor health.

- Compare the proportion of low APGAR scores for preterm births (less than 37 weeks gestation) with the corresponding proportion for full term births. Construct a confidence interval for the difference.

	Full term	Preterm	Total
High APGAR	268	93	361
Low APGAR	14	25	39
Total	282	118	400

- Parameters:
  p_1 = proportion with low APGAR for population of preterm births
  p_2 = proportion with low APGAR for population of full term births.
  We want a 95% confidence interval for $p_1 - p_2$.

- Verify independence and sample size conditions:
  - Randomization condition: We have a simple random sample from each population. Or a randomized experiment with two treatments.
  - 10% condition: 282/3065 and 118/1222.
  - The two samples are independent.
  - Success/Failure: Expected number of successes and failures in each sample is at least 10.
  - The data here are a SRS of size 400 from a Royal Alexandra Hospital database. If we condition on total counts for preterm and full term births, then that in effect yields two independent SRSs.

- Statistics: $\hat{p}_1 = \frac{y_1}{n_1}$ and $\hat{p}_2 = \frac{y_2}{n_2}$

- The sampling distribution model for $\hat{p}_1 - \hat{p}_2$ is Normal with

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \quad \text{and} \quad SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- Why this formula for SD? Samples are independent, so

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2).$$

- Why a Normal model? Because the difference of two independent Normal random variables is Normal.

- Standard error: $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

- Critical value: $z^* = 1.96$ for 95% confidence.
- Interval: $$(\hat{p}_1 - \hat{p}_2) \pm z^* \, SE(\hat{p}_1 - \hat{p}_2)$$
- Calculations:

$$\begin{aligned} \hat{p}_1 &= 25/118 = 0.2119 \\ \hat{p}_2 &= 14/282 = 0.0496 \\ SE &= \sqrt{0.001415 + 0.000167} = 0.0398 \\ (0.2119 - 0.0496) &\pm (1.96)(0.0398) \\ 0.162 &\pm 0.078 \end{aligned}$$

- Conclusion: I am 95% confident that the proportion of low 5-minute APGAR scores for preterm births in the Royal Alex database (target population) is between 8.4 and 24.0 percentage points higher than the corresponding proportion for full term births.

- Example. In 2005, the polling firm Ipsos-Reid was hired by Lavalife (dating website) to ask Canadian adults if they would date a smoker. They found that 56% of the 884 polled said they would not. In 2008, another poll was carried out asking the same question. This time 47% of 6313 said they would not date a smoker. Is this decrease statistically significant?

- Hypotheses: What are the target populations?

  $p_1 =$ proportion of 2005 population who would have said that they would not date a smoker. $p_2 =$ corresponding proportion for 2008 population.

  Test $H_0 : p_1 - p_2 = 0$ versus $H_A : p_1 - p_2 \neq 0$.

- Model: Verify conditions: randomization, 10%, independent groups, success/failure. Recall the sampling distribution model for $\hat{p}_1 - \hat{p}_2$.

  How would we estimate $SD(\hat{p}_1 - \hat{p}_2)$ if we knew H_0 was true?

Use a pooled estimator of the common proportion in the SE formula:

$$\hat{p}_1 = \frac{y_1}{n_1} \quad , \quad \hat{p}_2 = \frac{y_2}{n_2} \quad , \quad \hat{p}_{\text{pooled}} = \frac{y_1 + y_2}{n_1 + n_2}$$

$$SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}}$$

Test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)}$$

The sampling distribution model for z is standard Normal.

Compare the observed value of z with the standard Normal model and calculate the two-tailed P-value.

- Mechanics.

	2005	2008	Pooled
$y =$ number who say "no"	495	2967	3462
$n =$ sample size	884	6313	7197
$\hat{p} =$ sample proportion	0.560	0.470	0.481

$\hat{p}_1 - \hat{p}_2 = 0.090$, $SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = 0.0179$,

$z = 5.03$, two-sided P-value < 0.0001.

- Conclusion: The very small P-value provides clear-cut evidence that the proportion of the 2008 population not be opposed to dating a smoker is less than the corresponding proportion for the 2005 population.

- Is there evidence that proportions of newborns with low APGAR scores differs for males and females? Data from Royal Alexandra Hospital.

	Female	Male
Sample size	2062	2223
Low APGAR count	113	120
Sample proportion	0.0548	0.0540

- A US national poll on September 23, 2008, showed the following support for the Democratic and Republican presidential candidates:
  - Obama at $47\% \pm 2\%$, McKean at $44\% \pm 2\%$.
  - Is the difference statistically significant?
  - How would you obtain a P-value.

- "Own a pet and you'll likely use the net" (Edmonton Journal, 2000).
  - A random telephone survey of 400 Alberta households found that 69.0% use the internet.
  - The percentage was 76.5% among those who own a cat or dog.
  - The percentage was 66.0% among those who do not own a cat or dog.
  - Is there evidence that pet ownership was related to internet use?

	Pets	No pets	Pooled
Sample size	n_1	n_2	400
Number of internet users	$n_1\hat{p}_1$	$n_2\hat{p}_2$	
Proportion of internet users	$\hat{p}_1 = 0.765$	$\hat{p}_2 = 0.660$	0.690

- Example: Canada/USA health crisis?
  - Several years ago, an Environics poll of Canadians and Americans asked "Is your country's health-care system 'in a state of crisis' or 'basically in good shape'?"
  - Consider the proportion in each country who would have answered 'crisis' if they had been asked. Is there evidence that these two proportions are different?
  - Does the observed difference have practical significance?

	Canada	USA	Pooled
Sample size	3221	1008	
Number of 'crisis' responses	2158	635	
Proportion			