Last few slides from last time…

Binomial distribution

- Distribution of number of successes in n independent trials.
- Probability of success on any given trial = p
- Probability of failure on any given trial = q = 1-p

Mean and variance of a binomial random variable

- The mean number of successes in a binomial experiment is given by:
  - $\mu = np$
  - n is the number of trials, p is the probability of success
- The variance is given by
  - $\sigma^2 = npq$
  - $q = 1-p$

25 coin flips

- What is the probability that the number of heads is ≤ 14?

- We can calculate from the binomial formula that $p(x \leq 14)$ is .7878 (note this is *not* an approximation)

Normal Approximation

- Using the normal approximation with

  $\mu = np = (25)(.5) = 12.5$ and

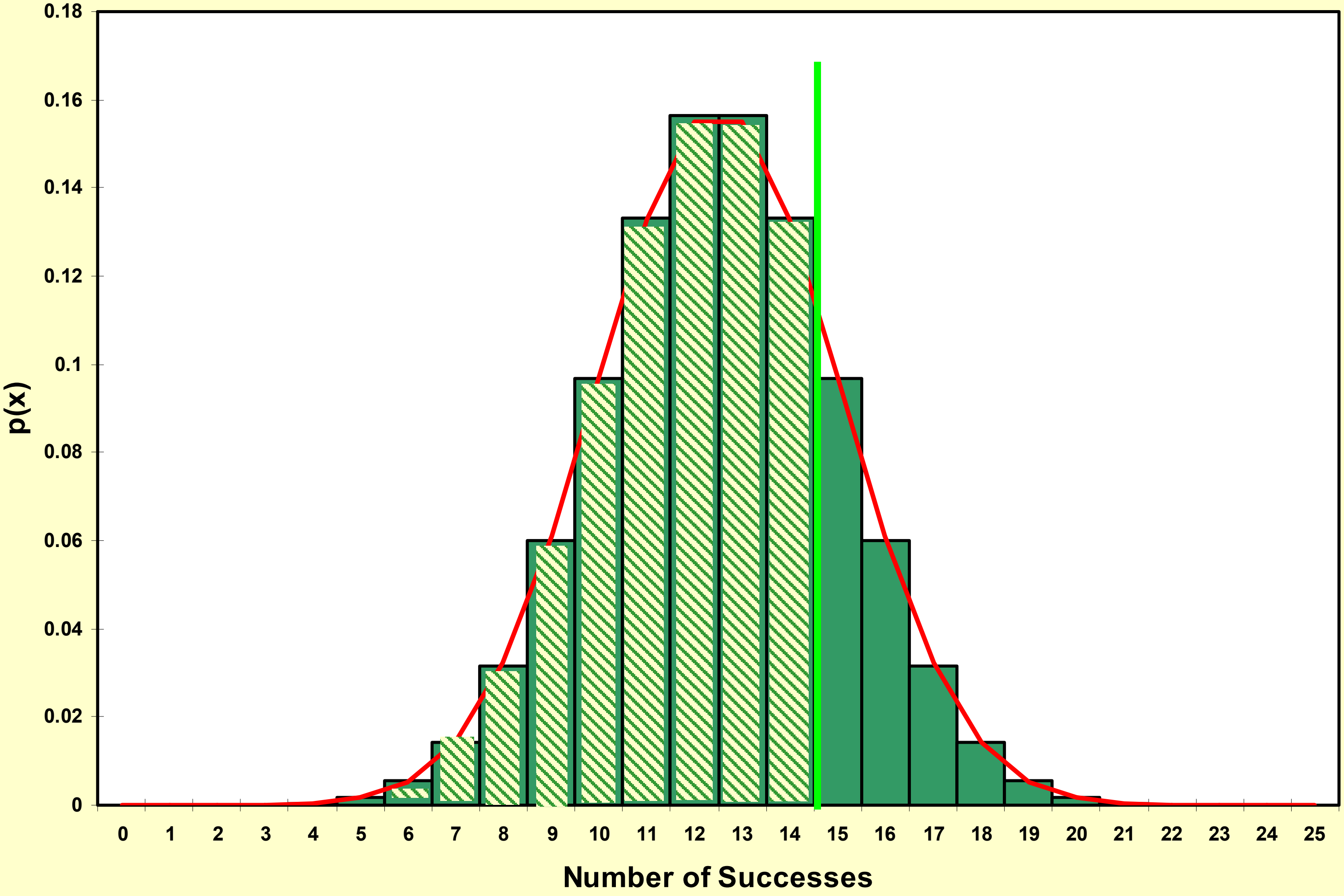
  $\sigma = \text{sqrt}(npq) = \text{sqrt}((25)(.5)(.5)) = 2.5$ we get

- $p(x \leq 14) = p(z \leq (14-12.5)/2.5))$

  $= p(z \leq .6) = .7257$

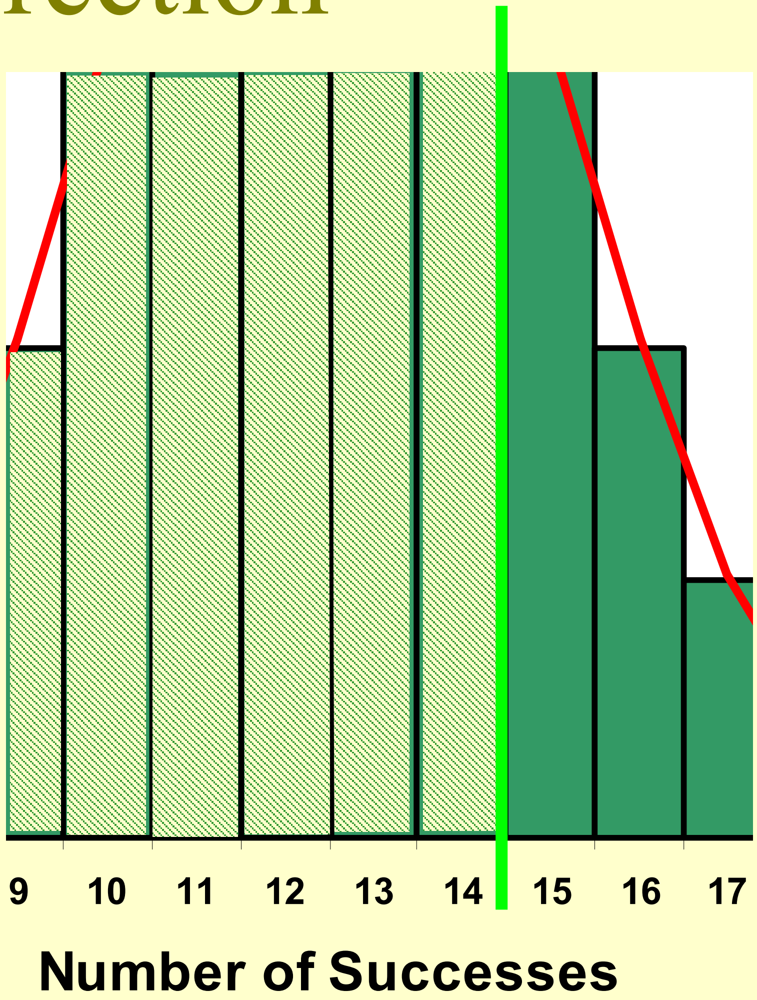
- .7878 vs. .7257 -- not great!!
- Need a better approximation...

Normal Approximation of Binomial Distribution



Continuity Correction

- Notice that the bars are centered on the numbers
- This means that $p(x \leq 14)$ is actually the area under the bars less than $x=14.5$
- We need to account for the extra 0.5
- $P(x \leq 14.5) = p(z \leq .8) = .7881$ -- a much better approximation!



Sampling theory

9.07

2/24/2004

Goal for rest of today

- *Parameters* are characteristics of *populations*.
  - E.G. mean, variance
- We've also looked at *statistics*[*] of a *sample*.
  - E.G. sample mean, sample variance
- How good are our statistics at estimating the parameters of the underlying population?

[*] *Statistic* = a function of a sample. There are a lot of possible statistics, but some are more useful than others.

Experimental design issues in sampling

- Suppose we want to try to predict the results of an election by taking a survey.
  - Don't want to ask *everyone* – that would be prohibitive!
  - What percent of voters will vote for the Republican candidate for president?
  - We ask 1000 eligible voters who they will vote for.
  - In the next lecture, we will talk about estimating the population parameter (the % of voters who prefer the Republican) from the % of the survey respondents who say they favor him.
  - But our statistics are only as good as our sampling technique – how do we pick those 1000 people for the survey?

Simple random sample

- If the procedure for selecting n objects out of a large population of objects is such that all possible samples of n objects are equally likely, then we call the procedure a *simple random sample*.
- This is the gold standard for sampling.
  - Unbiased: each unit has the same probability of being chosen.
  - Independent: selection of one unit has no influence on selection of other units.

In theory, how to get a simple random sample

- Get a list of every unit in the population.
- Randomly pick n objects using, e.g. a random number generator.

- Or, put a card for each unit in a drum, and pull out n cards (without looking)

- This may be prohibitive…

Opportunity sampling

- Take the first n people who volunteer
- The population of people who volunteer may be quite different from the general population.
  - Shere Hite: 100000 questionaires for book “Women & Love” left lying about in women’s organizations. Women returned the questionaires if they wanted to.
    - Came under fire because women in women’s organizations who volunteer to fill out the survey may have very different attitudes toward sex and love than the general population of women.

Opportunity sampling

- Nonetheless, in human studies, we do a lot of opportunity sampling in BCS.
- We're hoping there's not much difference between people who sign up for a cog sci experiment and the general population.
- In many cases this is probably not a bad assumption, but beware!
- Volunteering aside, many of the subjects are MIT students…

Simple random sample

- The methods described in this class apply to a simple random sample. If you don’t have one, the methods need to be modified.
- Randomized design is key to trusting your statistics.

Sampling theory

- A statistic can be computed from a sample, and used to estimate a parameter (of the population).
- A statistic is what the investigator knows.
- A parameter is what the investigator wants to know.
- When estimating a parameter, one major issue is accuracy: how close is the estimate going to be.

Freedman et al, Statistics

How good is our estimator?

- As an example, consider estimating the population mean, μ, with the mean of N samples, m.
- Bias:
  - If E(m) = μ, the estimator is *unbiased.*
  - If E(m) = μ', the bias is μ'- μ
  - All else being equal, you'd prefer that your estimate of the mean would, on average, equal the population mean, instead of, e.g., being smaller than μ, on average.

How good is our estimator?

- Consistency:
  - If the estimator gets better as we apply it to a larger sample, then the estimate is *consistent*.
- Relative efficiency:
  - Just as estimators have a mean, they also have a variance.
  - If G and H are both unbiased estimators of μ, then the more *efficient* estimator is the one with the smaller variance.
  - Efficient estimators are nice because they give you less chance error in your estimation of the population parameter.

How good is our estimator?

- Sufficiency
  - If an estimator G contains all of the information in the data about parameter μ, then G is a sufficient estimator, or *sufficient statistic*. That is, if G is a sufficient statistic, we can’t get a better estimate of μ by considering some aspect of the data not already accounted for in G.

What do we need, to decide how good our estimator is?

- Well, we need to know its mean and variance, so we can judge its bias and efficiency.
- It'd also be nice to know, more generally, what is the distribution of values we expect to get out of our estimator, for a given set of population parameters.
- Example: what is the distribution of the sample mean, given that the population has mean μ and variance σ^2 ?
  - This is called the *sampling distribution of the mean*.
  - You already estimated it on one of your homeworks.

“Sampling distribution” of the mean

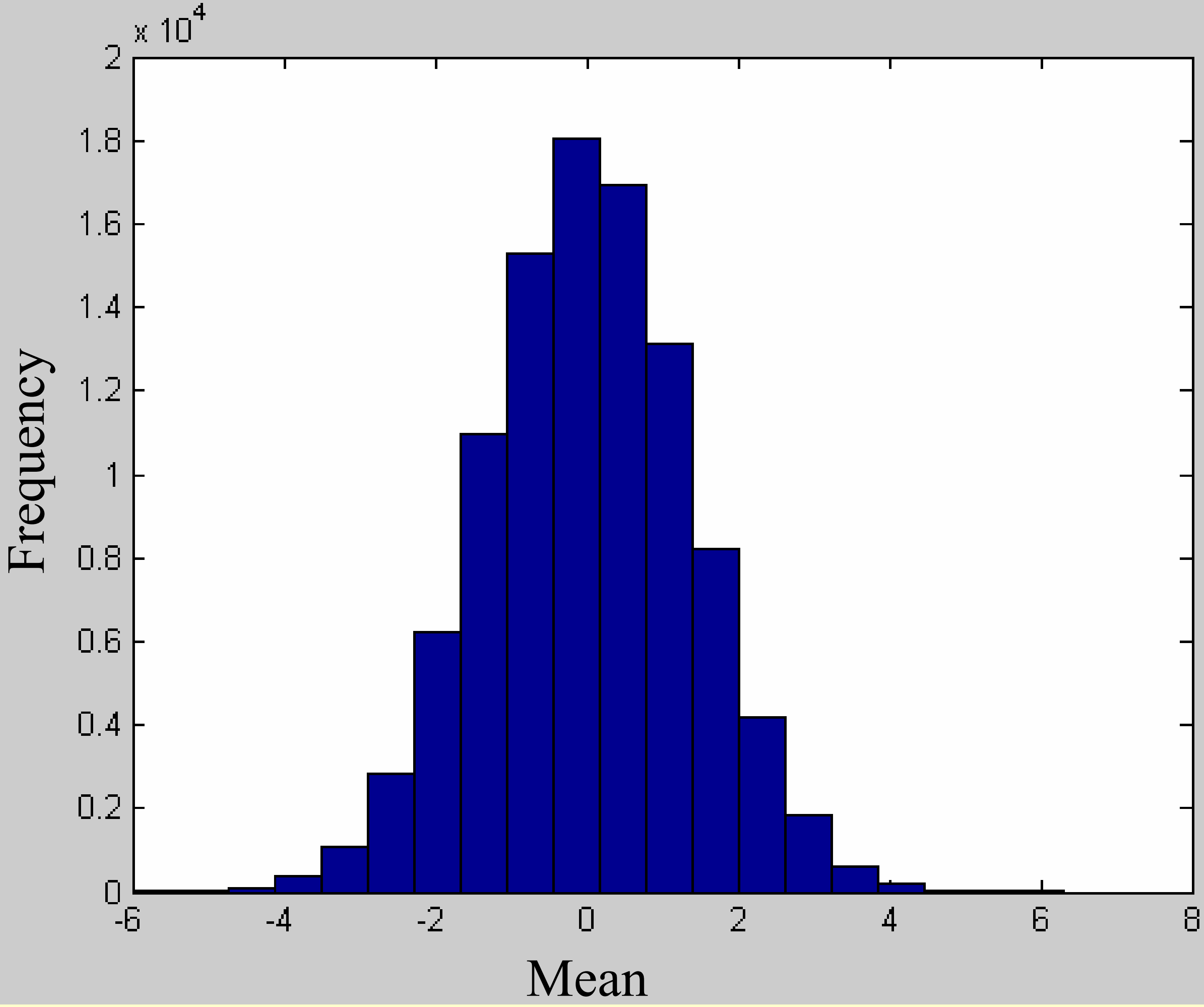
- On the last homework, you generated 100000 examples of 5 samples from a normal distribution with mean 0 and standard deviation 3.

```
x = randn(5, 100000)*3;
```

- You then took the mean of each of the 5 samples.

```
m = mean(x);        % mean of each of the columns
```

- Finally, you plotted a histogram of these means, m.


x 10^4
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Frequency
-6
-4
-2
0
2
4
6
8
Mean

“Sampling distribution” of the mean

- The histogram of the means, m, was your estimate of the *sampling distribution* of the mean.
- A sampling distribution is a distribution of what happens when you take N samples from some distribution, and compute some statistic of those samples.
  - In this case, N=5, “some distribution” = *N*(0,3), and “compute some statistic” = compute the mean.
- A sampling distribution is a theoretical probability distribution. You were estimating it by running the experiment (take 5 samples, take the mean) a large number of times (100000), and looking at the histogram of the results.

Questions

- From your estimate of the sampling distribution of the mean, what seems to be the shape of this distribution?
- Its mean, E(m)?

The distribution of the sample mean, m = the sample distribution of the mean

- What is its mean?
- Its standard deviation?
- Its shape?

But, first, let's look at another one of your homework problems

- You generated a bunch of samples of $x \sim N(\mu_x, \sigma_x^2)$, $y \sim N(\mu_y, \sigma_y^2)$, then created a new random variable $z = x + y$.
- What, did you estimate, is μ_z?
  - It turns out this is true, in general.
- What, did you estimate, is σ_z^2?
  - It turns out this is true for any x, y independent.

Facts: If x and y are independent

- $z = x+y \rightarrow \sigma_z^2 = \sigma_x^2 + \sigma_y^2$
- Of course, this generalizes to

  $z=x_1+x_2+\ldots+x_N$

  $\sigma_z^2 = \sigma_{x1}^2 + \sigma_{x2}^2 + \ldots + \sigma_{xN}^2$
- Note also that $z = x-y \rightarrow \sigma_z^2 = \sigma_x^2 + \sigma_y^2$

Facts for *any* random variables x and y

- E(x + y) = E(x) + E(y)
  - Intuition: if you play two games, x and y, the amount you expect to win is the sum of the amount you expect to win in game x, and the amount you expect to win in game y.
  - DON'T CONFUSE THIS WITH p(x+y), WHICH IS NOT NECESSARILY EQUAL TO p(x) + p(y)!!!!!!!
- E(kx) = k E(x)

So, what is E(m)?

- Let x_i be distributed according to some distribution with mean μ and variance σ^2.
- $m = (x_1+x_2+\ldots+x_N)/N$
- $E(m) = E[(x_1+x_2+\ldots+x_N)/N]$
  $= 1/N\ [E(x_1)+E(x_2)+\ldots+E(x_N)]$
  $= 1/N\ [N\ \mu]$
  $= \mu$
- $E(m) = E(\text{mean of N samples } x_i) = E(x) = \mu$

The sample mean is an unbiased estimator for μ

- E(m) = E(mean of N samples x_i) = E(x) = μ
- You should have suspected this from your homework (though of course your *estimate* of E(m) differed by a small amount from μ).

What is the standard deviation of m?

- Recall $x \sim \mathcal{N}(\mu, \sigma^2)$
- $m = (x_1+x_2+\ldots+x_N)/N,$
- We know that if $z=x_1+x_2+\ldots+x_N$
  $\sigma_z{}^2 = \sigma_{x1}{}^2 + \sigma_{x2}{}^2 + \ldots + \sigma_{xN}{}^2$
  $= N\,\sigma^2$
- $m = z/N$
- Recall from an earlier lecture that if $y=kx$,
  $\sigma_y = k\sigma_x$ -> $\sigma_y{}^2 = k^2\,\sigma_x{}^2$
- So, $\sigma_m{}^2 = N\,\sigma^2/N^2 = \sigma^2/N$
  $\sigma_m = \sigma/\mathrm{sqrt}(N)$

The standard error

- To distinguish this standard deviation of the mean from the standard deviation of x, the standard deviation of the mean is known as the *standard error*.
  - Standard error of the mean = σ/sqrt(N)
- The sample mean is relatively *efficient*, when compared to the median (see HW due next week).
- The sample mean is also *consistent* – the standard error will go down as we increase the number of samples, and our estimate will get better.

What is the shape of the sampling distribution of the mean, m?

- If x is normal, the distribution of the mean is normal.
- If x is not normal, don’t despair – the central limit theorem says that for N large enough (and certain other constraints, like finite σ^2), the distribution of the mean is approximately normal.

Example 1

- x approximately normal, with mean = 20, standard deviation 4.
- Experiment: you observe 50 samples from x. What is the probability that their mean lies between 19 and 21?
- $E(m) = 20$
- Std. error of the mean = $\sigma_m = 4/\mathrm{sqrt}(50)$
  ≈ 0.5657
- $z(21) = (21-E(m))/\sigma_m \approx 1.77$; $z(19) \approx -1.77$
- From z-tables, $P(-1.77 < z < 1.77) \approx 92.4\%$
  $= P(19 < m < 21)$
- Compare with $P(19 < x < 21) \approx 19.7\%$

Example 2

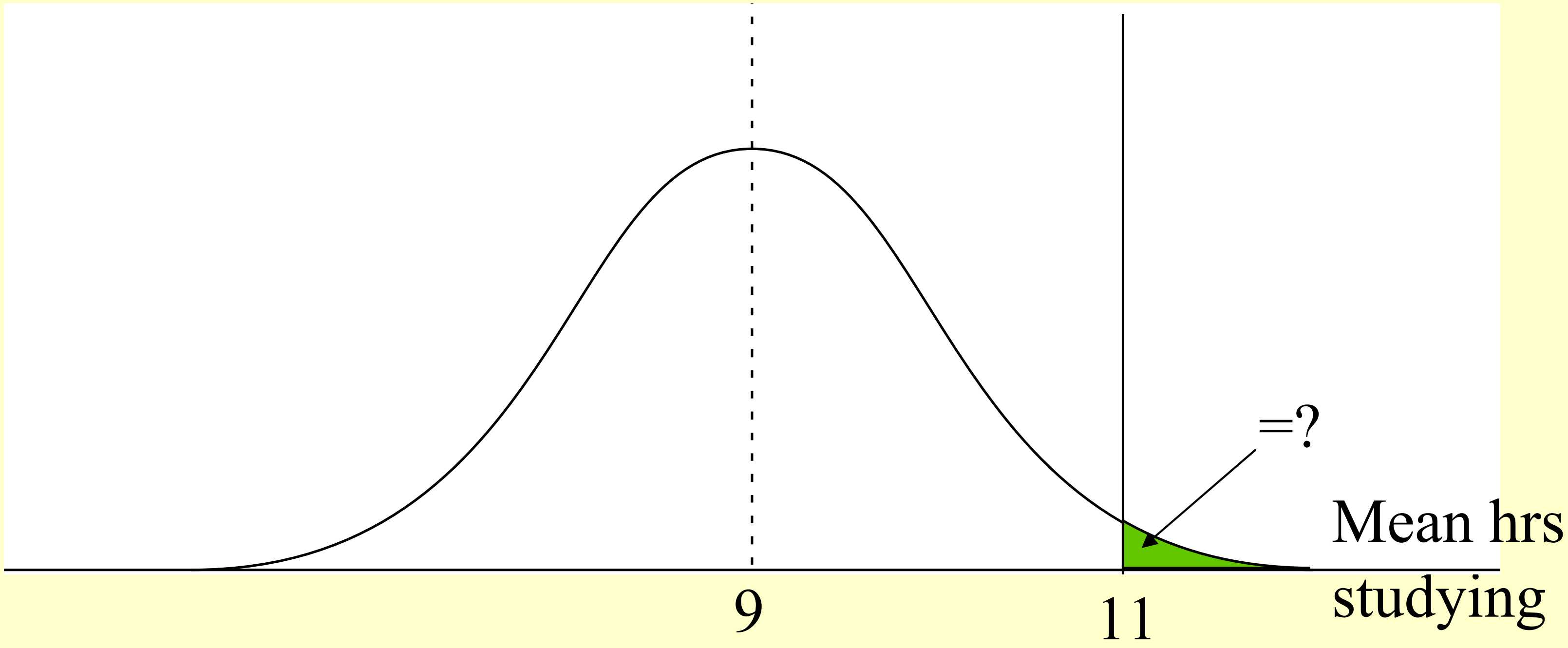
- What if we had been computing the sample mean from only 30 samples?
- E(m) = 20, as before
- Std. error of the mean = σ_m = 4/sqrt(30) $\approx$ 0.73
- z(21) = (21-E(m))/$\sigma_m \approx$ 1.37; z(19) $\approx$ -1.37
- From z-tables, P(-1.37 < z < 1.37) $\approx$ 82.9%
  - If you average over only 30 samples, the estimate of the mean gets considerably more variable.

Example 3: What is the probability that the sample mean will fall in a certain range, given the population mean?

- You survey 25 students taking a class with 3-0-9 credit units. For such a class, they are supposed to spend about 9 hours in homework and preparation for the class per week. The students spend on average 11 hours studying for this class, with a standard deviation of 4 hours.
- If the true population mean is 9 hours of studying per week, what is the probability that we would have seen a sample mean of as high as 11 hours per week?

Probability of observing a sample mean of 11 hrs/week

- We want to know: what is the probability that the sample mean for our sample of 25 students is 11 hours or more, given that the true mean is 9 hours.



Probability and the sample mean

- 11 hrs is how many std. errors away from the mean of 9 hrs?
- Std. error of the mean =
  std. deviation of the population/sqrt(N)
  - Don't know std. deviation of the population. Estimate that the population variance = the sample variance.
  - Std. error of the mean ≈ 4/sqrt(25) = 4/5
- z = (11-9)/(4/5) = 2.5
- p(z > 2.5) = (1-0.9876)/2 = 0.0062
- It seems highly unlikely that we would see a mean as high as 11 in our sample of 20 students, given a true mean of 9 hrs. Probably this class is too much work for 3-0-9 credit hours.

Problems with using the normal approximation to estimate probabilities

- In order to use the normal approximation, we need to know σ, the standard deviation of the underlying population. Often we don't, as in the previous example.
  - But, we can estimate σ from the standard deviation of the sample.
- The distribution of the sample mean may only be normal for large sample sizes, N.

When you don't know σ

- But, z = (m-μ)/(σ/sqrt(N)) is distributed according to a z distribution (*N*(0, 1))
- Whereas t = (m-μ)/(s/sqrt(N)) is distributed according to a *t-distribution*
  - Invented by William Gosset at Guinness brewery early in the 20th century
  - Called the Student's t distribution, because he published it anonymously

Creating the t-distribution

- You could try doing it in MATLAB
  - Generate an Nx 1000000 array of samples from a normal distribution with some mean and variance.
  - Look at the distribution of (m-μ)/(s/sqrt(N))
- Gosset created the t-tables without MATLAB (or modern computers) (ugh!)
- Note that there is a different t-distribution for each value of N (the number of samples you use you compute the sample mean).

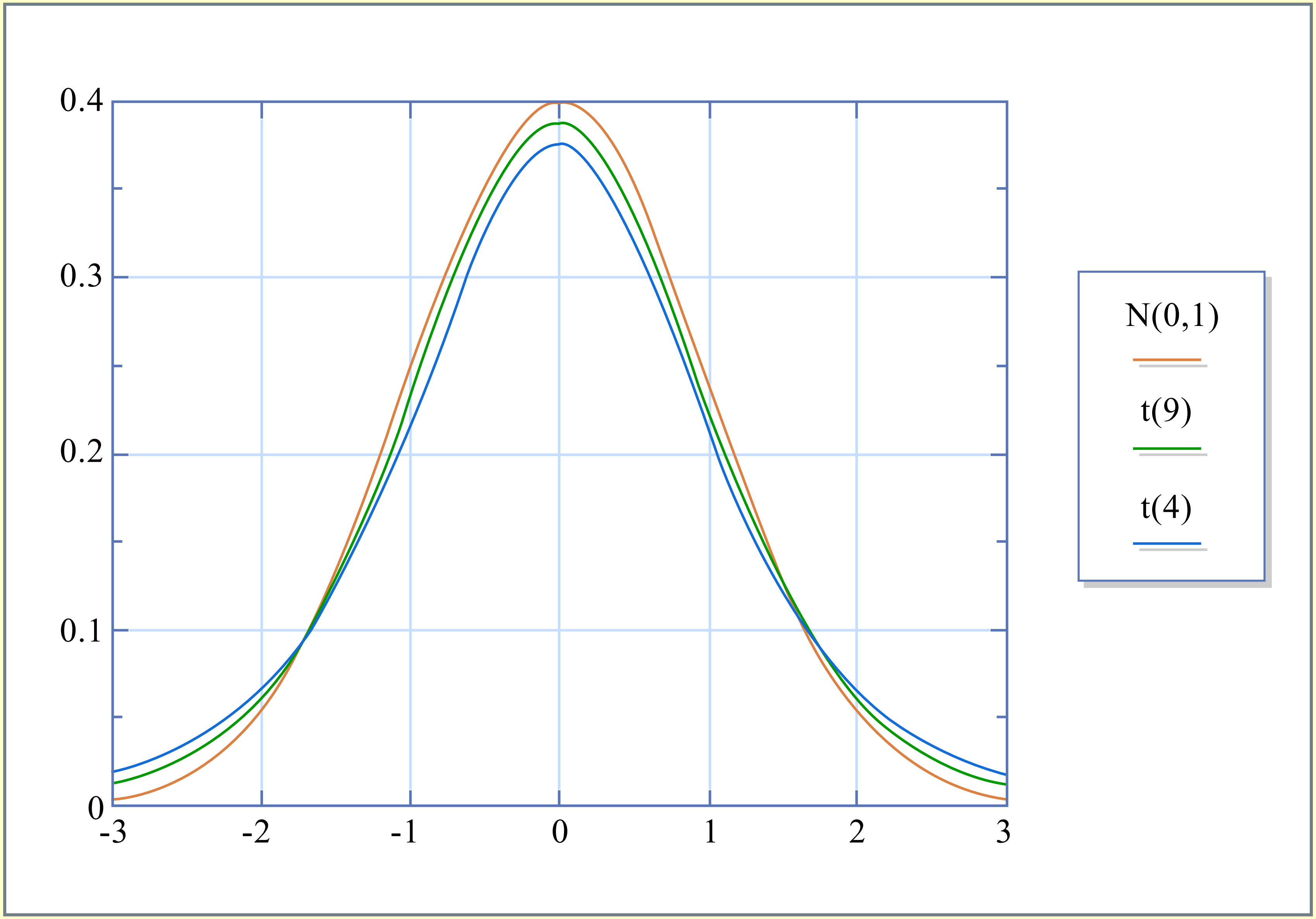


Figure by MIT OCW.

Student’s t Distribution

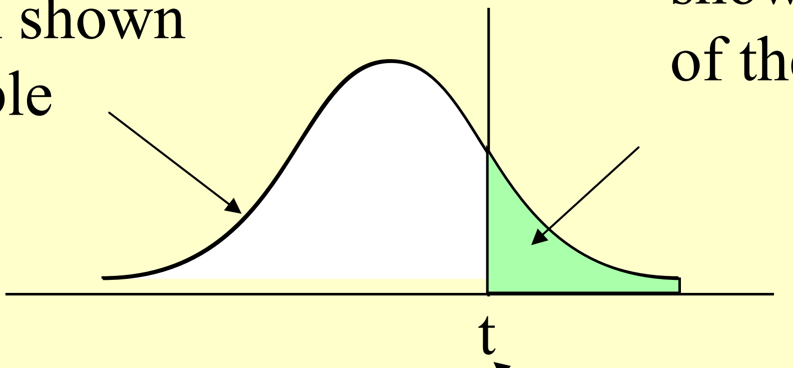
- The distribution of t is more spread out than the z distribution. It is not normal!

- Using s instead of σ introduces more uncertainty, making t “sloppier” than z

- As the sample size gets larger, s gets closer to σ, so t gets closer to z.

The Student's t-distribution

- t-table in the back of your book (A-106)



Student's curve, with degrees of freedom shown at the left of the table

The shaded area is shown along the top of the table

is shown in the body of the table

Degrees of freedom?

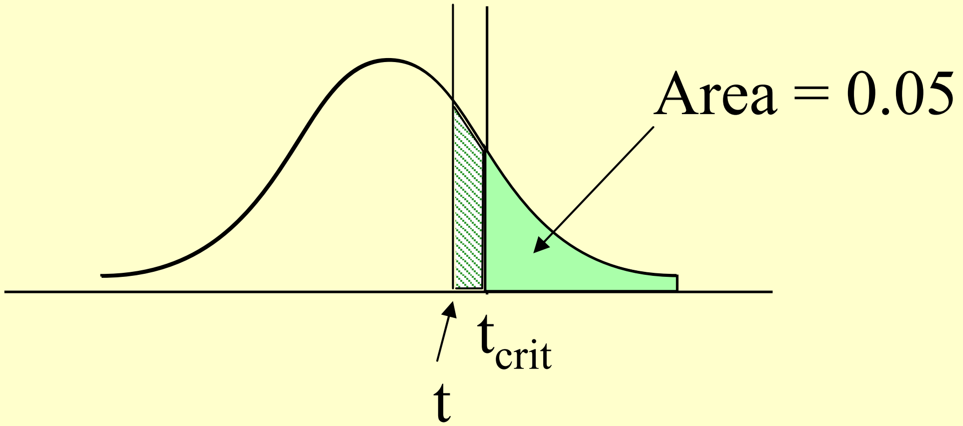
- If n is the sample size, n-1 is the “degrees of freedom” of the sample.
- The general idea: Given n pieces of data, you start out with n degrees of freedom. However, when you compute the mean, you are left with only n-1 degrees of freedom, because the mean constrains the n samples.
  - Given the mean and n-1 samples, you can figure out the nth sample.

Example of using the t-table

- You survey 5 students taking a class with 3-0-9 credit units. For such a class, they are supposed to spend about 9 hours in homework and preparation for the class per week. The students spend on average 11 hours studying for this class, with a standard deviation of 4 hours.
- If the true population mean is 9 hours of studying per week, is the probability that we would have seen a sample mean of as high as 11 hours per week < 5%?

Is P(m>11) < 0.05?

- t = (m-μ)/(s/sqrt(N)) = (11-9)/(4/sqrt(5)) ≈ 1.12
- What is t_{crit} for area=0.05?
- $t_{crit}(4) = 2.13$
- $t < t_{crit}(4)$, so P(m>11) > 0.05



- From these 5 samples, it seems feasible that the 11 hr average occurred by chance.

Statistical tables online

- By the way, a lot of statistical tables are available online, e.g:
- http://math.uc.edu/~brycw/classes/148/tables.htm

Problems with using the normal approximation to estimate probabilities

- In order to use the normal approximation, we need to know σ, the standard deviation of the underlying population. Often we don't, as in the previous example.
  - Use the t-distribution, with the sample estimate of σ.
- The distribution of the sample mean may only be normal for large sample sizes, N.
  - The t-distribution was created for the mean of normally distributed data, x, but it turns out to be pretty robust to non-normal x, so long as x is kinda "mound-shaped".

Other sample distributions

- This was all for the sample distribution of the mean
- Can also look at the sample distributions for other statistics, e.g. sample std. deviation, sample variance, median, …
- An important one: estimating the parameter of a binomial distribution

Sample distribution of proportions

- Binomial distribution, probability of success = p
- Want to estimate p.
- Candidate estimator: P' = sample proportion of successes on n trials.
  - What is mean(P')?
  - Std. deviation(P') = std. error?
  - Shape of the distribution?

Mean and std. deviation of a binomial distribution

- Recall for binomial distributions with probability of success = p, probability of failure = 1-p = q
  - – Mean = np
  - – Std. deviation = sqrt(npq)

Sampling distribution of proportions

- P' = number of successful trials/n
- Shape of the sampling distribution of P':
  - Approximately normal, if n was large enough that the distribution of the number of successful trials was approximately normal.
  - (Just divided every value by n – distribution still normal.)
- Mean(P') = np/n = p   Unbiased estimator of p.
- Std. deviation(P') = sqrt(npq)/n = sqrt(pq/n) = *standard error* of the proportion.

Binomial distribution, percents, and surveys

- “But wait, the binomial distribution required that p be the same on every trial. This meant we had to sample *with replacement*.
- If I call a bunch of people to ask them which politician they favor, that’s not sampling with replacement – if I call one person to ask their preference, I check them off my list and don’t call them again to ask the same question.”
- It’s true, this is sampling without replacement.
- It makes little difference that this is sampling without replacement, in the case of most surveys (for which the population is large relative to the sample size), but one can correct for this, if desired.
  - Before calling person i, p(Kerry) = 0.45 = 450000/1000000
  - After calling person i, p(Kerry) = 449999/999999 ≈ 0.45

Sampling without replacement

- Note same sampling without replacement comment is also valid for the earlier problem on number of hours spent studying.
- If the class (your population) only has 40 students, and you’ve asked 25 of them how much they studied, you need to adjust your estimate of the std. error.

Sampling without replacement

- This issue of sampling with vs. without replacement is a real issue in certain circumstances.
- However, it is rarely an issue for the sorts of work we do in BCS.
  - Our sample sizes are quite small compared to the size of the desired population
  - Total population size = # of humans, e.g.
- But, there is a correction, if you need it (E.G. sample size ~ ½ population size). Look in a stat book.

Example 1: Sampling distribution of the proportion

- 5 balls in a box. 1 red, 4 green. ($p = 0.2$)
- Experiment: you sample with replacement 50 times.
- What is the probability that p' lies between 0.13 and 0.27? Use the normal approximation to the binomial distribution.
- $E(P') = p = 0.20$
- Std. error of the proportion $= \sigma_{P'} = \mathrm{sqrt}(pq/n)$
  ≈ 0.0566
- $z(0.27) = (0.27-0.20)/\sigma_{P'} \approx 1.24$; $z(0.13) \approx -1.24$
- From z-tables, $P(-1.24 < z < 1.24) \approx 78.9\%$
  $\approx P(0.13 < p' < 0.27)$

Example 2

- What if we tried to estimate p using only 30 trials?
- E(P’) = p = 0.20
- Std. error of the proportion = $\sigma_{P'}$ = sqrt(pq/n)
  ≈ 0.0730
- z(0.27) = (0.27-0.20)/$\sigma_{P'}$ ≈ 0.96;
  z(0.13) ≈ -0.96
- From z-tables, P(-0.96 < z < 0.96) ≈ 66.3%
  ≈ P(0.13 < p’ < 0.27)
- Again, the estimate of p gets more variable as you look at fewer samples.

Example 3: What is the probability that p' will fall in a certain range, given p?

- Flip a coin 50 times. If the coin is fair (p=0.5), what is the probability of getting an estimate, p', greater than or equal to 0.7 (=35 heads).
- E(P') = 0.5
- Std. error(P') = sqrt((.5)(.5)/50) = .0707
- z = (0.7-0.5)/0.0707 ≈ 2.83
- P(z > 2.83) ≈ (1-0.9953)/2 = 0.0024 = P(p' > 0.7)
  - Pretty unlikely to get such high estimates of p

More examples on finding the mean and standard deviation of a r.v.

- $x \sim N(\mu_x, \sigma_x)$, $y \sim N(\mu_y, \sigma_y)$

- $Z = x + 4y + 2$
  - $E(Z) = E(x) + E(4y) + E(2) = \mu_x + 4\mu_y + 2$
  - $Var(Z) = var(x) + var(4y) + var(2)$
    $= \sigma_x^2 + 16\,\sigma_y^2$

- $Z = (2x_1 + 2x_2 - y)/5$
  - $E(Z) = (E(2x) + E(2x) - E(y))/5 = 4/5\,\mu_x - 1/5\,\mu_y$
  - $Var(Z) = var(2x/5) + var(2x/5) + var(y)$
    $= 8/25\,\sigma_x^2 + \sigma_y^2$