UCLA STAT 13

**Introduction to Statistical Methods for the Life and Health Sciences**

**Instructor: Ivo Dinov,**

Asst. Prof. of Statistics and Neurology

**Teaching Assistants:**

Fred Phoa, Kirsten Johnson, Ming Zheng & Matilda Hsieh

University of California, Los Angeles, Fall 2005

*http://www.stat.ucla.edu/~dinov/courses_students.html*



## Chapter 5

## Sampling Distributions



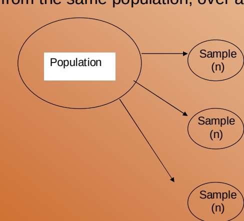
## Sampling Distributions

- **Definition:** *Sampling Variability* is the variability among random samples from the same population.
- A probability distribution that characterizes some aspect of sampling variability is called a sampling distribution.
  - tells us how close the resemblance between the sample and the population is likely to be.
- We typically construct a sampling distribution for a statistic.
  - Every statistics has a sampling distribution.



## The Meta-Experiment

- All the possible samples that might be drawn from the population (infinity repetitions).
  - In other words if we were to repeatedly take samples of the same size from the same population, over and over.






## The Meta-Experiment

- Meta-experiments are important because probability can be interpreted as the long run relative frequency of the occurrence of an event.
- Meta-experiments also let us visualize sampling distributions.
  - and therefore understand the variability among the many random samples of a meta-experiment.



## Dichotomous Observations

- Dichotomous - two outcomes
  - (yes or no, good or evil, etc…)
- We use the following notation for a dichotomous outcome
  - P population proportion
  - $\hat{p}$ sample proportion
- The big question is how close is $\hat{p}$ to P?
- To determine this we need to examine the sampling distribution of $\hat{p}$
- What we want to know is:
  - if we took many samples of size n and observed $\hat{p}$ each time, how would those values of be distributed around p?

## Dichotomous Observations

**Example**: Suppose we would like to estimate the true proportion of male students at UCLA. We could take a random sample of 50 students and calculate the sample proportion of males.

- What is the correct notation for:
  - the true proportion of males?
  - the sample proportion of males?
- Suppose we repeat the experiment over and over. Would we get the same proportion of males for the second sample?



## Reece's Pieces Experiment

**Example**: Suppose we would like to estimate the true proportion of orange reece's pieces in a bag. To investigate we will take a random sample of 10 reece's pieces and count the number of orange. Next we will make an approximation to a sampling distribution with our class results.

What you need to calculate:
- the number of orange
- the sample proportion of orange (number of orange/10)



## An Application of a Sampling Distribution

***Example***: Mendel's pea experiment. Suppose a tall offspring is the event of interest and that the true proportion of tall peas (based on a 3:1 phenotypic ratio) is 3/4 or p = 0.75. If we were to randomly select samples with n = 10 and p = 0.75 we could create a probability distribution as follows:

Lab_Mendel_Pea_Experiment.html
(work out in discussion/lab)

Validate using:
http://socr.stat.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm
E.g., B(n=10, p=0.75, a=6, b=6)=0.146

| $\hat{p}$ | Number Tall | Number Dwarf | Probability |
|---|---|---|---|
| 0.0 | 0 | 10 | 0.000 |
| 0.1 | 1 | 9 | 0.000 |
| 0.2 | 2 | 8 | 0.000 |
| 0.3 | 3 | 7 | 0.003 |
| 0.4 | 4 | 6 | 0.016 |
| 0.5 | 5 | 5 | 0.058 |
| 0.6 | 6 | 4 | 0.146 |
| 0.7 | 7 | 3 | 0.250 |
| 0.8 | 8 | 2 | 0.282 |
| 0.9 | 9 | 1 | 0.188 |
| 1.0 | 10 | 0 | 0.056 |



## An Application of a Sampling Distribution

- What is the probability that 5 are tall and 5 are dwarf?

P(5 tall and 5 dwarf) = P( $\hat{p}$ = 5/10)

= P( $\hat{p}$ = 0.5)

= 0.058

| $\hat{p}$ | Number Tall | Number Dwarf | Probability |
|---|---|---|---|
| 0.0 | 0 | 10 | 0.000 |
| 0.1 | 1 | 9 | 0.000 |
| 0.2 | 2 | 8 | 0.000 |
| 0.3 | 3 | 7 | 0.003 |
| 0.4 | 4 | 6 | 0.016 |
| 0.5 | 5 | 5 | 0.058 |
| 0.6 | 6 | 4 | 0.146 |
| 0.7 | 7 | 3 | 0.250 |
| 0.8 | 8 | 2 | 0.282 |
| 0.9 | 9 | 1 | 0.188 |
| 1.0 | 10 | 0 | 0.056 |



## An Application of a Sampling Distribution

- If we think about this in terms of a meta-experiment and we sample 10 offspring over and over, about 5.8% of the $\hat{p}$'s will be 0.5.
  - This is the sampling distribution of sample proportion of tall offspring is the distribution of in repeated samples of size 10.
- If we take a random sample of size 10, what is the probability that six or more offspring are tall?

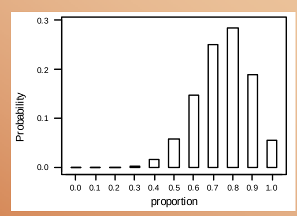
P($\hat{p}$ $\geq$ 0.6) = 0.146 + 0.250 + 0.282 + 0.188 + 0.056

= 0.922



## An Application of a Sampling Distribution

- This table could also be represented as a histogram with probability on the y-axis and proportion on the x-axis.
  - easier to draw these by hand



## Relationship to Statistical Inference

- We can also use our sampling distribution of  to estimate how much sampling error there is within 5 percentage points of p. Because we knew p from the previous example (p=0.75), we might want to estimate:

  $P(0.7 \le \hat{p} \le 0.8)$
  = 0.250 + 0.282 = 0.532

  There is a 53% chance that for a sample of size 10, $\hat{p}$ will be within ± 0.05 of p.

  - This seems a little crazy, why?

| $\hat{p}$ | Number Tall | Number Dwarf | Probability |
|---|---|---|---|
| 0.0 | 0 | 10 | 0.000 |
| 0.1 | 1 | 9 | 0.000 |
| 0.2 | 2 | 8 | 0.000 |
| 0.3 | 3 | 7 | 0.003 |
| 0.4 | 4 | 6 | 0.016 |
| 0.5 | 5 | 5 | 0.058 |
| 0.6 | 6 | 4 | 0.146 |
| 0.7 | 7 | 3 | 0.250 |
| 0.8 | 8 | 2 | 0.282 |
| 0.9 | 9 | 1 | 0.188 |
| 1.0 | 10 | 0 | 0.056 |



## Relationship to Statistical Inference

- So far we have been using p to determine the sampling distribution of $\hat{p}$.
- Why sample for $\hat{p}$ when we already know p?
  - We don't need to know p to get a good estimate (this will come later).



## Sample Size

- As n gets larger, $\hat{p}$ will become a better estimate of p.
- Just to show…

| N | $P(0.7 \le \hat{p} \le 0.8)$ |
|---|---|
| 10 | 0.53 |
| 20 | 0.56 |
| 50 | 0.673 |
| 100 | 0.798 |

*These calculations were done using the SOCR binomial distribution Calculator.
http://socr.stat.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm

E.g., B(n=20, p=0.75, a=0.7x20=14, b=0.8x20=16)=0.5606

THE POINT: A larger sample improves the chance that $\hat{p}$ is close to p.

- Caution: this doesn't necessarily mean that the estimate will be closer to p, only that there is a better chance that it will be close to p.