Sampling Distributions

- Most of our histograms show distributions of scores
- But sample statistics (e.g., mean) also have distributions
  - these are called sampling distributions
  - “higher-level” distributions of statistics, not scores
  - variation across samples, not individuals
  - every statistic (mean, standard deviation, r) has a sampling distribution
- Sampling distributions are IMPORTANT
  - used to make inferences about population parameters

Example: Using Sampling Distributions of r

Histograms show distributions of r for scrambled (X,Y) data

Red dashed lines indicate boundaries which contain 95% of scrambled r’s

r’s outside boundaries are rare/unusual

Observed r’s fall outside 95% boundaries, so they are unusual if we assume that the null hypothesis (pop r = 0) is true
Sampling Distributions & Inference

- sampling distributions are linked to statements about probability
- we need to know something about their shapes
- usually we do not know the shape of distributions of scores
  - so how can we know anything about shape of distribution of sample statistics like the mean?
- we use the **Central Limit Theorem** ...

Sampling Distribution of the Mean

- Randomly select set of \( n \) scores \( X \) drawn from population
  - scores are distributed in population:
    - mean = \( u \), variance = \( \sigma^2 \), shape of distribution = UNKNOWN
- Compute mean of sample & record for later use
- Repeat first 2 steps many times
- How is sample mean distributed? i.e., **What is sampling distribution of mean?**
  - In the long run, average mean will equal population mean (\( u \))
  - In the long run, variance equals population variance (\( \sigma^2 \)) divided by \( n \)
- **Central Limit Theorem:**
  - if sample \( n \) is sufficiently large, the sampling distribution of the mean will be a normal distribution, regardless of the shape of the distribution of scores

Central Limit Theorem

If a random variable \( Y \) has a population mean \( u \) and a population variance \( \sigma^2 \), then the sample mean, \( M \), based on \( n \) observations, has an approximate normal distribution with mean \( u \) and a variance \( \sigma^2/n \) for sufficiently large \( n \).

---

The Normal Distribution

- Normal distribution
  - “Bell Curve”; Gaussian distribution
- used often in statistics
- unimodal & symmetrical around mean
  - defined by mean and standard deviation

---

---
Some Normal Distributions

Why the Normal Distribution?

- Many variables we measure are normally distributed
- Many statistical procedures assume data are normally distributed
- Assuming a variable is normally distributed allows us to
  - describe the shape of the distribution
  - calculate probabilities for different values of variable
- **Central Limit Theorem** means that the assumption about normality often is reasonable

Central Limit Theorem Example

Uniform Distribution of Scores

Each sample consists of 4 scores selected randomly from uniform distribution ranging from -3.2 to +7.2. Means calculated for many samples (n=4). How are means distributed?
Sampling Distribution of the Mean is nearly Normal when $N=4$.

Sampling Distribution of the Mean is not Normal when $N=4$.

Sampling Distribution of the Mean is approximately Normal when $N=40$.

Log-Normal Distribution of Scores

Each sample consists of $N$ scores selected randomly from a log-normal distribution (positive skew). Means calculated for many samples. How are means distributed?
### z Transformation (Standard Scores)

- Normal “distribution” is a family of distributions $N(\mu, \sigma^2)$
- We can convert any normal distribution into a standard normal distribution by transforming the values of our $X$ into a standard normal score, $z$
- $z$ scores are distributed as $N(0,1)$
  - Normal, mean = 0, variance = 1

$$z = \frac{X - \mu}{\sigma}$$

### Example: z transform

1. Subtract the value of the mean from each score.
   - $50 - 50 = 0$
   - $40 - 50 = -10$
   - And so on...

2. Divide each variable by $\sigma$.
   - $0/10 = 0$
   - $-10/10 = -1$
   - $+10/10 = +1$
   - And so on...

### Z-scores

- If we transform our $X_i$ values into $Z_i$ values, these are called z-scores
- The value of $z$ represents how many standard deviations, $\sigma$, an observation, $i$, is from the mean, $\mu$
- Often use z-scores to compare distributions — similar to using $r$ to compare linear trends between different pairs of (X,Y) variables

### The standard normal distribution

- A normal distribution with mean ($\mu$) = 0, and variance ($\sigma^2$) = 1
- Denoted $N(0, 1)$, where N stands for normal, zero is mean, and 1 is variance
- Use distribution to calculate probabilities:
Areas under the curve represent probabilities

The area under the curve between points $Z_1$ and $Z_2$, or the integral, is equal to the probability that a random score will fall within that interval.

Finding area from a table

Instead of calculating the integral/area, we often use statistical tables.

Making use of z-scores

- What is the probability of randomly selecting a score that is at least 2.5 standard deviations higher than the mean?

### Table 6.1
The Normal Distribution

<table>
<thead>
<tr>
<th>$t$</th>
<th>Mean to $t$</th>
<th>Larger Portion</th>
<th>Smaller Portion</th>
<th>$t$</th>
<th>Mean to $t$</th>
<th>Larger Portion</th>
<th>Smaller Portion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.05</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.05</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.10</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.10</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.15</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.15</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.20</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.20</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.25</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.30</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.30</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.35</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.35</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.40</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.40</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.45</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.45</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.50</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.50</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.55</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.55</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.60</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.65</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.65</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.70</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.70</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.75</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.75</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.80</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.80</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.85</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.85</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.90</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.90</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.95</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
<tr>
<td>0.95</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>1.00</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.5000</td>
</tr>
</tbody>
</table>
We consult our z-table...

Q. Which of these values gives us the answer we're looking for?

Making use of z-scores

• The probability of selecting a score that is more than 2.5 standard deviations above the mean is 0.0062

Q1. What is the probability of selecting a score more than 2.5 SDs below the mean? p(z < -2.5) = 0.006

Q2. What is the probability of getting a score more than 2.5 SDs above or below the mean? p(z < -2.5) OR p(z > 2.5) = 0.006 + 0.006 = 0.012

Q3. What is the probability of randomly selecting a score between -1 and -2 SDs from the mean?
Q3. What is the probability that a student would score between -1 and -2 standard deviations from the mean?

We consult our z-table...

Setting probable limits on an observation

What is the z-score cutoff for the lower 85% of the scores?

We consult our z-table...
Using standardized scores (z test)

A woman in the US has just given birth to a full-term baby weighing 291 kg. Is this weight unusually low?

z test

- In US, population of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g.
- The weights are distributed approximately normally.
- A weight of 2910 g is 1.23 standard deviations below the mean:
  \[-z = (2910 - 3480) / 462 = -1.23\]
- What is the probability of observing a weight that is at least this low?
*z test*

- In US, population of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g.
- The weights are distributed approximately normally.
- A weight of 2910 g is 1.23 standard deviations below the mean:
  \[
  z = \frac{2910 - 3480}{462} = -1.23
  \]
- What is the probability of observing a weight that is at least this low?
  \[
  p(z < -1.23) = 0.109
  \]