4.7 Probability Plots

4.8 The Central Limit Theorem

4.7 Probability Plots

Check that data plausibly came from a probability distribution before using that distribution.

A *probability plot* graphs $\{(x_i, y_i)\}$, where $\{x_i\}$ are a ____________ random sample and $\{y_i\}$ are n _______________ from the distribution. The points should be __________________________ if the sample is from the distribution. Here's how to make a probability plot:

- Let $c_i = \frac{i-\frac{1}{2}}{n}$, the ____________ of the i^{th} of n equal-length subintervals of $[0, 1]$
- Let $y_i =$ the $(100c_i)^{\text{th}}$ ____________ of the distribution: $P(X < y_i) = c_i$
- Graph the pairs $\{(x_i, y_i)\}$ along with the line ____________.

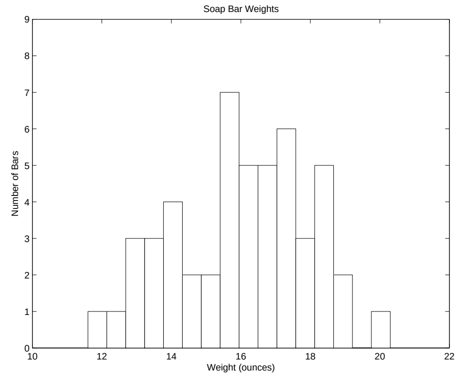
e.g. To check whether $\{x_1 = -1, x_2 = 0, x_3 = 2\}$ are from $N(0, 1)$ (usually ______ for a good plot),

- $c_1 =$ ____________, $c_2 =$ ____________, $c_3 =$ ____________
- $\{y_i\}$ are such that $P(Z < y_1) =$ ________, $P(Z < y_2) =$ ________, $P(Z < y_3) =$ ________
  $\implies y_1 =$ ________, $y_2 =$ ________, $y_3 =$ ________
- Graph (-1, ________), (0, ________), (2, ________) along with ____________

e.g. (p. 160 #2) Make a normal probability plot for the 50 soap bars weights in §1.3 #1 (p. 30). Do they appear to come from an approximately normal distribution?

- Here are the weights (ounces), $x_1, \cdots, x_{50}$, already sorted:

11.6	12.6	12.7	12.8	13.1	13.3	13.6	13.7	13.8	14.1
14.3	14.3	14.6	14.8	15.1	15.2	15.6	15.6	15.7	15.8
15.8	15.9	15.9	16.1	16.2	16.2	16.3	16.4	16.5	16.5
16.5	16.6	17.0	17.1	17.3	17.3	17.4	17.4	17.4	17.6
17.7	18.1	18.3	18.3	18.3	18.5	18.5	18.8	19.2	20.3



Since $\bar{x} = 16.03$ and $s_x = 1.95$, check whether the weights are plausibly from ___________________.

- Here are the centers $\{c_i\}$ of 50 subintervals of $[0, 1]$:

.01	.03	___	.07	.09	.11	.13	.15	.17	.19
.21	.23	.25	.27	.29	.31	.33	.35	.37	.39
.41	.43	.45	.47	.49	.51	.53	.55	.57	.59
.61	.63	.65	.67	.69	.71	.73	.75	.77	.79
.81	.83	.85	.87	.89	.91	.93	.95	.97	.99

- We would standardize y_i by $z_i = \frac{y_i - \mu}{\sigma} \approx \frac{y_i - \bar{x}}{s_x}$; z_i is the $(100c_i)^{\text{th}}$ percentile of $N(0,1)$:

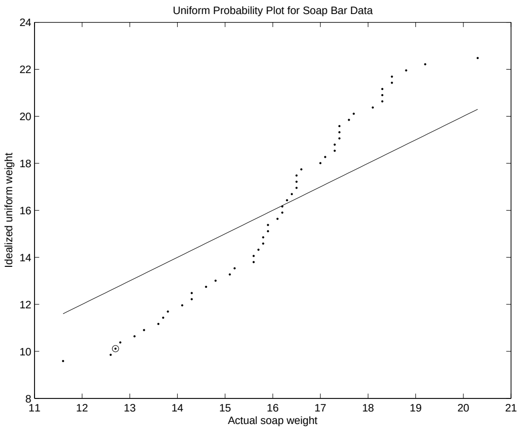
-2.33	-1.88	____	-1.48	-1.34	-1.23	-1.13	-1.04	-0.95	-0.88
-0.81	-0.74	-0.67	-0.61	-0.55	-0.50	-0.44	-0.39	-0.33	-0.28
-0.23	-0.18	-0.13	-0.08	-0.03	0.03	0.08	0.13	0.18	0.23
0.28	0.33	0.39	0.44	0.50	0.55	0.61	0.67	0.74	0.81
0.88	0.95	1.04	1.13	1.23	1.34	1.48	1.64	1.88	2.33

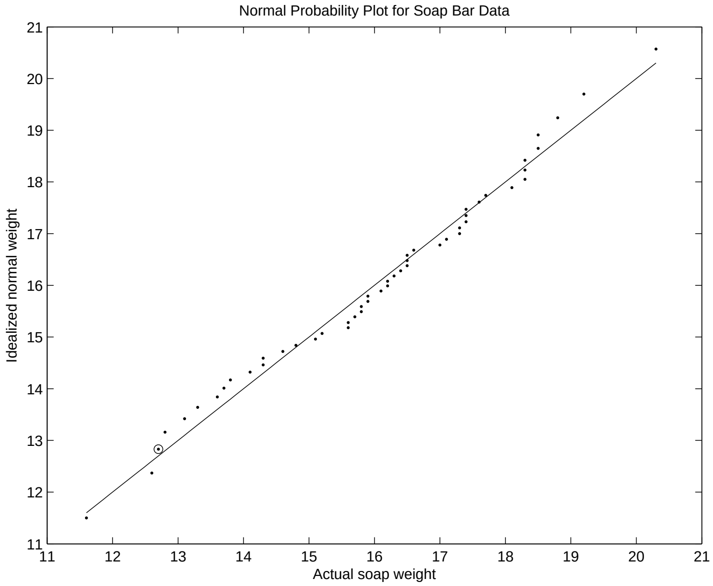
Unstandardize by computing $y_i = \bar{x} + z_i s_x$, the $(100c_i)^{\text{th}}$ percentile of $N(\bar{x}, s_x^2) = N(16.03, 1.95^2)$:

11.5	12.4	____	13.2	13.4	13.6	13.8	14.0	14.2	14.3
14.5	14.6	14.7	14.8	15.0	15.1	15.2	15.3	15.4	15.5
15.6	15.7	15.8	15.9	16.0	16.1	16.2	16.3	16.4	16.5
16.6	16.7	16.8	16.9	17.0	17.1	17.2	17.3	17.5	17.6
17.7	17.9	18.1	18.2	18.4	18.6	18.9	19.2	19.7	20.6

Finally, graph $\{(x_i, y_i)\}$ with the line $y = x$. Are the points nearly on a line? ____________ $\Longrightarrow$

____

- 



4.8 The Central Limit Theorem

The *Central Limit Theorem* (CLT) says that the mean, $\bar{X}$, of a large sample from (almost) ____________ distribution with finite μ and σ, is $\approx$ ____________:

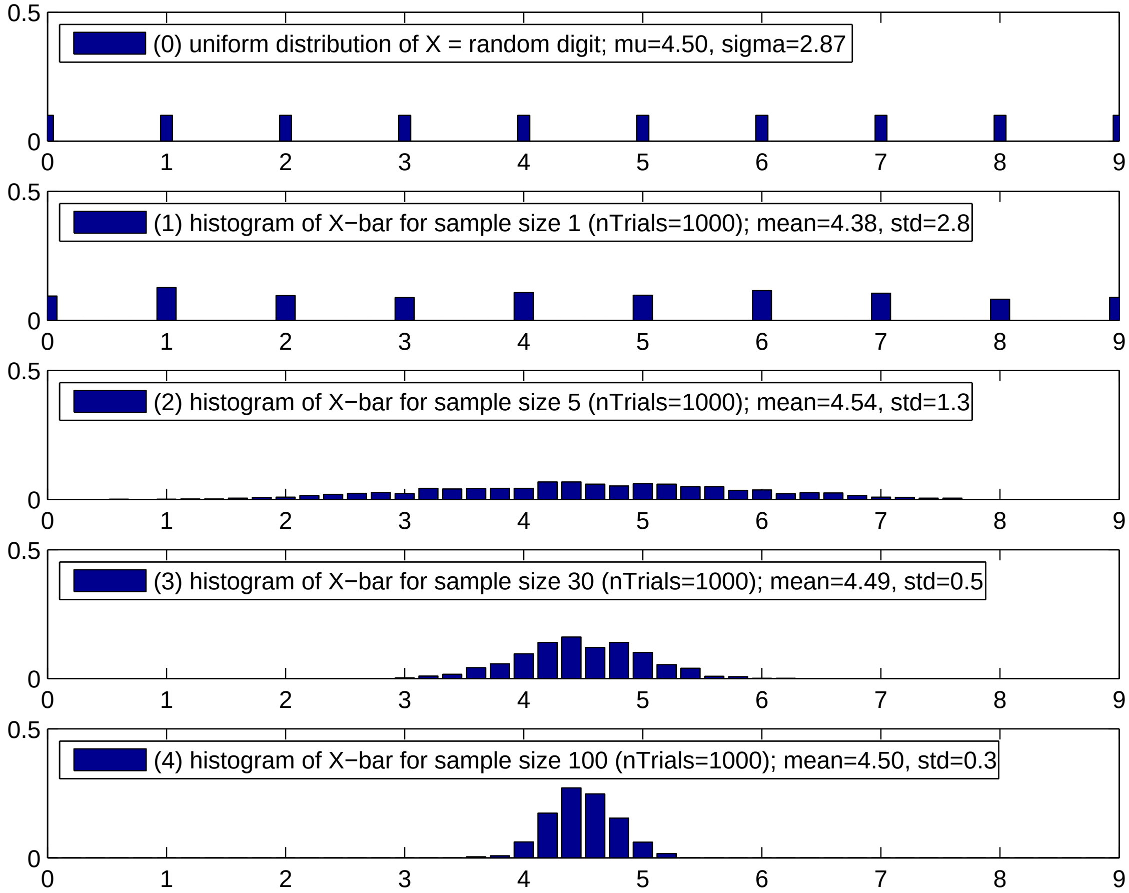
If $X_1, \cdots, X_n$ is a simple random sample from a population with mean μ and variance σ^2, and n is ____________________, then $\boxed{\bar{X} \sim N(\mu, \frac{\sigma^2}{n})}$ (approximately). Or, in terms of the sum of the data, $S_n = \sum X_i = n\bar{X} \sim N($________, ________$)$. ($n > 30$ often counts as "sufficiently large".)

Sir Francis Galton, who gave us ______________, ______________, and ____________, said:

I know of scarcely anything so apt to impress the imagination as the wonderful form

> of ________________ expressed by the [Central Limit Theorem]. The law would have been personified by the Greeks and _________, if they had known of it. It reigns with serenity ... amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a ________________ of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of _____________ proves to have been latent all along. (*Natural Inheritance*, 1889)

e.g. Here is a simulation of the generation of many random samples from the discrete distribution with mass function $p(x) = \frac{1}{10}$ for $x \in \{0, 1, \cdots, 9\}$ (and 0 otherwise):

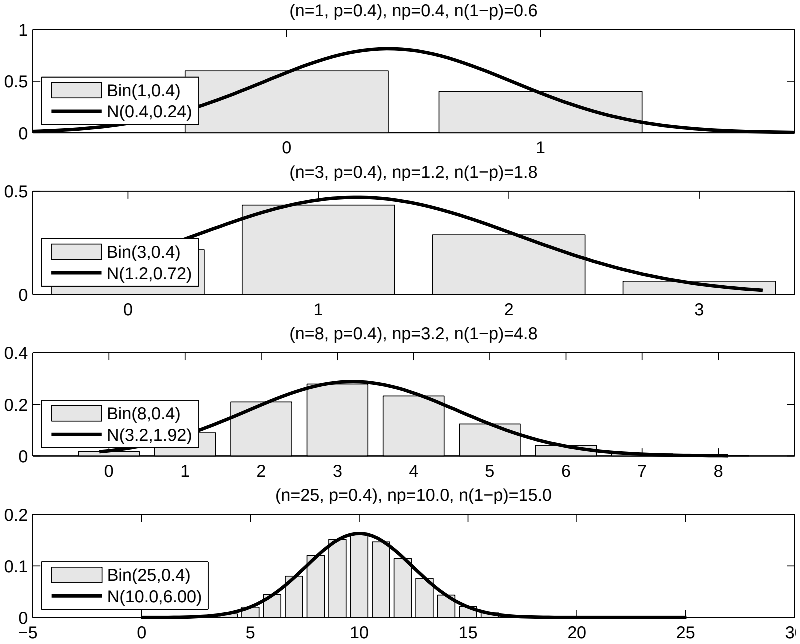


e.g. An insurance company knows that in the population of millions of homeowners, the mean annual loss from fire is $\mu = \$250$ and the standard deviation is $\sigma = \$1000$. (The loss distribution is strongly right-skewed, since most policies have no loss but a few have large losses.) If the company sells 10,000 policies, can it safely base its rates on the assumption that the average loss will be no greater than $275?

Normal Approximation to the Binomial

If $X \sim \text{Bin}(n, p)$, and $np > 10$ and $n(1-p) > 10$, then $X \sim N($________,__________$)$ (approximately; because the ______________ applies to $X = \sum_{i=1}^{n} Y_i$, where $Y_i \sim \text{Bernoulli}(p)$).

e.g. Here are graphs of $\text{Bin}(n, p)$ and $N(np, np(1-p))$ for several n and p:



A *continuity correction* addresses the fact that $X \sim \text{Bin}(n, p)$ is ______________, while $Y \sim N(np, np(1-p))$ is ______________, by moving the endpoint at which Y is evaluated. For an integer x (#successes), $P(X < x)$ is usually closer to $P(Y < x - \frac{1}{2})$ than to $P(Y < x)$. (Similarly, $P(X \leq x)$ is closer to $P(Y < x + \frac{1}{2})$ than to $P(Y < x)$.)

Normal Approximation to the Poisson

Since $\text{Bin}(n, p) \approx N(np, np(1-p))$ for $np > 10$ and $n(1-p) > 10$ (above); and (§4.2) for large n and small p, if $\lambda = np$, then $\text{Bin}(n, p) \approx \text{Poisson}(np)$; it follows that, for $\lambda > 10$, $\text{Poisson}(\lambda) \approx N($_____,_____$)$.

e.g. (of CLT; p. 168 #10) A battery manufacturer claims that the lifetime of a type of battery has $\mu = 40$ hours and $\sigma = 5$ hours. Suppose a random sample of 100 batteries is selected.

(a) If the claim is true, what is $\text{P}(\bar{X} \leq 36.7)$? ______________. (b) Based on (a), if the claim is true, is $\bar{X} = 36.7$ unusually short? ______________. (c) If $\bar{X} = 36.7$, is the claim plausible? ________.

(d) If the claim is true, what is $P(\bar{X} \leq 39.8)$? ______________. (e) Based on (d), if the claim is true, is $\bar{X} = 39.8$ unusually short? ______________. (f) If $\bar{X} = 39.8$, is the claim plausible? ________.